THE INTUITIVE APPEAL OF EXPLAINABLE MACHINES

Andrew D. Selbst[*]
Solon Barocas[†]

As algorithmic decision-making has become synonymous with inexplicable decision-making, we have become obsessed with opening the black box. This Article responds to a growing chorus of legal scholars and policymakers demanding explainable machines. Their instinct makes sense; what is unexplainable is usually unaccountable. But the calls for explanation are a reaction to two distinct but often conflated properties of machine-learning models: inscrutability and non-intuitiveness. Inscrutability makes one unable to fully grasp the model, while non-intuitiveness means one cannot understand why the model's rules are what they are. Solving inscrutability alone will not resolve law and policy concerns; accountability relates not merely to how models work, but whether they are justified.

In this Article, we first explain what makes models inscrutable as a technical matter. We then explore two important examples of existing regulation-by-explanation and techniques within machine learning for explaining inscrutable decisions. We show that while these techniques might allow machine learning to comply with existing laws, compliance will rarely be enough to assess whether decision-making rests on a justifiable basis.

[*] Postdoctoral Scholar, Data & Society Research Institute; Visiting Fellow, Yale Information Society Project. Selbst is grateful for the support of the NSF 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 Kaminsky, Rónán Kennedy, Been Kim, Jon Kleinberg, Brian Kreiswirth, Chandler May, Brent Mittlestadt, Deidre Mulligan, David Lehr, Paul Ohm, Helen Nissenbaum, Frank Pasquale, 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.

We argue that calls for explainable machines have failed to recognize the connection between intuition and evaluation and the limitations of such an approach. A belief in the value of explanation for justification assumes that if only a model is explained, problems will reveal themselves intuitively. Machine learning, however, can uncover relationships that are both non-intuitive and legitimate, frustrating this mode of normative assessment. If justification requires understanding why the model's rules are what they are, we should seek explanations of the process behind a model's development and use, not just explanations of the model itself. This Article illuminates the explanation-intuition dynamic and offers documentation as an alternative approach to evaluating machine learning models.

INTRODUCTION

There can be no total understanding and no absolutely reliable test of understanding.

—Joseph Weizenbaum, *Contextual Understanding by Computers* (1967)[1]

Complex, inscrutable algorithms increasingly inform consequential decisions about all 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 could be unnerving,[3] unfair,[4] unsafe,[5] unpredictable,[6] and unaccountable.[7] How can inscrutable

[1] 10 COMM. ACM 474, 476 (1967). In the 1960's, the project of 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] 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/; Aaron M Bornstein, *Is Artificial Intelligence Permanently Inscrutable?*, NAUTILUS (Sept. 1, 2016), http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable;

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

[4] *See, e.g.,* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF L. REV. 671, 677-692 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2018).

[5] *See, e.g.,* David Lazer, et al., *The Parable of Google Flu: Traps in Big Data Analysis*, 343 SCI. 1203 (2014).

[6] *See, e.g.*, Jamie Condliffe, *Algorithms Probably Caused a Flash Crash of the British Pound,* MIT TECHNOLOGY REV. (Oct. 7, 2016), https://www.technologyreview.com/s/602586/algorithms-probably-caused-a-flash-crash-of-the-british-pound/; Curtis E.A. Karnow, *The Application of Traditional Tort Theory to*

algorithms be held to account for bad 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. People have a natural feel for explanation. We know how to offer explanations, and can often agree when one is good, bad, or in between, when the explanation is on point or off-topic. Lawyers in particular 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.[8]

As comfortably as we use explanations, however, asking someone to define the concept will more often than not generate a blank look in response. Analytically, explanation is infinitely variable, and there can be many valid explanations for a given phenomenon or decision. For example, a partial list of reasons for a glass having shattered include: a) because it hit the ground; b) because it was dropped; c) because the holder was startled (and that's why it was dropped); d) because gravity pulled it toward the earth; e) because glass is brittle; f) because the ground is solid (and therefore harder than brittle glass); g) because of the chemical composition of glass (making it brittle), and so on. These are all valid explanations, some nested within others, and some having nothing to do with each other. How should we choose what type of explanation to offer?

Thus far, the scholarly discourse in both law and machine learning around explanation has primarily revolved around similar questions—which kinds of explanations are most useful and which are technically available? But 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 anti-

Embodied Machine Intelligence, *in* ROBOT LAW 51, 57-58 (Ryan Calo, A. Michael Froomkin & Ian Kerr, eds. 2016) (discussing unpredictability in robotics).

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

[8] Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633, 636-37 (1995); *see also* TOM R. TYLER, WHY PEOPLE OBEY THE LAW (1990).

discrimination or due process. It is therefore important to ask how exactly we engage with those machine explanations in order to connect them to the normative questions of interest to law.

If someone offers an "explanation" of why a glass shattered, it will be immediately obvious to the questioner whether that explanation responds to the concern at hand (e.g., her broken barware or a question in chemistry class). As if by miracle, humans make these inferences easily. Even better, when they fail to do so, the back-and-forth of human communication permits correction (e.g., the chemistry teacher pointing out that she is actually asking about the shattered beaker in the lab, not the chemical composition of glass). But when asked to describe this process of inference, it is quite difficult. When humans process information without being able to rationally engage with it, we usually call it intuition.[9]

In this Article, we argue that scholars and advocates who seek to use explanation to enable justification of machine learning models are relying centrally on intuition. Intuition is both powerful and highly flawed. We argue that while this mode of justifying decision-making remains important, we must understand the benefits and weaknesses of connecting machine explanation to intuitions. To remedy the limitations of intuition, we must also consider alternatives, which include institutional processes, documentation, and access to those documents.

The Article proceeds in four Parts. Part I examines the various anxieties surrounding the use and justification of automated decision-making. After discussing secrecy, a lack of transparency, and a lack of technical expertise, we argue that the two concepts that truly set machine learning decision-making apart are inscrutability and non-intuitiveness. These concepts are similar, but distinct and easily confused.

Part II examines laws and machine learning tools designed specifically to explain inscrutable decisions. On the legal side, the Part discusses the "adverse action notices" required by federal credit laws and the informational requirements of Europe's new General Data Protection Regulation. On the technical side, the Part discusses various techniques used by computer scientists to make machine learning models interpretable, including keeping them simple by design, global rule extraction, tools to

[9] Merriam-Webster, *Intuition*, https://www.merriam-webster.com/dictionary/intuition (2.c: "the power or faculty of attaining to direct knowledge or cognition without evident rational thought and inference").

extract the most important factors in a particular decision, and interactive methods. The Part evaluates the limitations of the focus on the black box in both law and technology.

Part III builds up to the connection between explanation and intuition before evaluating the merits of an intuition-centered approach to justification. It begins by canvassing the reasons besides justification that one might want interpretable machines: explanation as an inherent good and explanation as a way to enable action in data subjects or consumers. Neither is adequate to fully address the concerns with automated decision-making. Interrogating the assumptions behind a third reason—that explanation will reveal flawed or acceptable bases for decision-making—demonstrates the reliance on intuition. The remainder of the Part examines the upsides and downsides of intuition. With respect to machine learning in particular, while intuition will be able to root out obviously good or bad cases, it will not capture the cases that give machine learning its greatest value: true patterns that exceed human imagination. These are not obviously right or wrong, but simply strange.

Part IV aims to provide another way. Once we leave the black box, we are left to question the humans involved. There are large parts of the process of machine learning that do not show up in the model but can contextualize the result: choices made and not made, costs associated with better models, etc. If we cannot intuitively explain models, sometimes justification can be achieved by demonstrating due care and thoughtfulness. Such demonstrations can be achieved through the existence of institutional processes and documentation, coupled with access to those documents, which can either be public by design (impact statements) or open to oversight based on some trigger (litigation). There will still be hard cases, but documentation will allow strange cases to become less strange, and should aid to our typical intuition-driven modes of justification when they fall down in the face of machine learning systems.

Machine learning models are not magic.[10] They can be broken down and examined. Where intuition fails in any other domain, the answer is to become more rational, more scientific—to examine further. Just so here. Where machine interpretability fails to engage intuition, the answer is

[10] Madeleine C. Elish & danah boyd, *Situating Methods in the Magic of Big Data and Artificial Intelligence*, COMM'N MONOGRAPHS 57, 62–63 (2017); Andrew D. Selbst, *A Mild Defense of Our New Machine Overlords*, 70 VAND. L. REV. EN BANC 87, 104 (2017).

to examine more. To do so, the law needs to demand the processes and documentation necessary to permit that examination.

I. INSCRUTABLE AND NON-INTUITIVE

Scholarly and policy debates over how to regulate a world controlled by algorithms have been mired in difficult questions about how to observe, access, audit, or understand those algorithms.[11] The difficulty of regulating algorithms has been attributed to a diverse set of problems, specifically that they are "secret"[12] and "opaque"[13] "black boxes"[14] that are rarely if ever made "transparent";[15] that they operate on the basis of correlations rather than "causality"[16] and produce "predictions"[17] rather than "explanations"[18]; that their behavior may lack "intelligibility"[19] and

[11] Solon Barocas, Sophie Hood & Malte Ziewitz, *Governing Algorithms: A Provocation Piece*, (2013); Malte Ziewitz, *Governing Algorithms: Myth, Mess, and Methods*, 41 SCI. TECH. & HUMAN VALUES 3 (2016); Nick Seaver, *Knowing Algorithms*, PROC. MEDIA IN TRANSITION 8 (unpublished manuscript), *available at* https://static1.squarespace.com/static/55eb004ee4b0518639d59d9b/t/55ece1bfe4b030b2e8302e1e/1441587647177/seaverMiT8.pdf; Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO. COMM'N & SOC'Y 14 (2017).

[12] *See, e.g.*, 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, *passim* (2011); Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. TELECOMM. & HIGH TECH. L. 235, 237 (2011).

[13] Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC. 1, 3–5, Jan.–Jun. 2016; 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, *passim*.

[14] *See. e.g.*, FRANK PASQUALE, The BLACK BOX SOCIETY (2015).

[15] *See, e.g.*, Citron & Pasquale, *supra* note 7, *passim*; Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503 (2013).

[16] *See, e.g.*, Kim, *supra* note 4, at 875.

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

[18] *See, e.g.*, Bryce Goodman & Seth Flaxman, *European Union Regulations On Algorithmic Decision-Making And A "Right To Explanation"* 38 AI MAGAZINE 50 (2016).

[19] *See, e.g.*, Brennan-Marquez, *supra* note 17, at 1253.

"foreseeability;"[20] and that they challenge established ways of "being informed"[21] or "knowing."[22] These terms are frequently used interchangeably or assumed to have overlapping meanings. For example, opacity is often seen as a synonym for secrecy,[23] an antonym for transparency,[24] and, by implication, an impediment to understanding.[25] And yet the perceived equivalence of these terms has obscured important differences between distinct problems that frustrate attempts at regulating algorithms—problems that require disentangling before the question of regulation can even be addressed.

In this Part, we argue that many of the above challenges, while important, are not unique to algorithms or to machine learning. We seek here to parse the problems raised by machine learning more precisely, arguing that rather than a lack of transparency, causality, or knowledge, the two properties that set machine learning apart from prior decision mechanisms are *inscrutability* and *non-intuitiveness*. We adapt and extend a taxonomy first proposed by Jenna Burrell,[26] where our primary purpose is to emphasize these last two properties and clear up confusion.[27] Inscrutability and non-intuitiveness have been conflated in the past: where the property of inscrutability suggests that fully transparent models may defy understanding, non-intuitiveness suggests that even where models are understandable, they may rest on relationships that defy intuition.

A. Secrecy

The first common critique of algorithmic decision-making is secrecy,

[20] *See, e.g.*, Karnow, *supra* note 6, *passim*.

[21] *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).

[22] Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, NEW MEDIA & SOCIETY 1, 2–5 (2016).

[23] Burrell, *supra* note 13, at 3–4.

[24] Zarsky, *supra* note 13, at 124; Ford & Price, *supra* note 13, at 12.

[25] Burrell, *supra* note 13, at 4–5.

[26] *See generally* Burrell, *supra* note 13.

[27] Our parsing of the different issues related to explanation is similar to the taxonomy proposed by Ed Felten on the Freedom to Tinker Blog. 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/.

the worry that the decision-making process may be completely hidden from those affected by it. This worry is as old as the original Code of Fair Information Practices (FIPs), the conceptual basis for the vast majority of privacy law.[28] The very first FIP is that "[t]here must be no personal-data record-keeping systems whose very existence is secret."[29] 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.[30] Such worries are particularly pronounced when the government engages in algorithmic decision-making,[31] but similar objections arise in the commercial sector, where there are a remarkable number of scoring systems of which consumers are simply unaware.[32] In many cases, this ignorance exists because the companies engaged in such scoring are serving other businesses rather than consumers.[33] 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. Notably, this objection does not speak to how these decisions are made.

[28] WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES (2018).

[29] ROBERT GELLMAN, FAIR INFORMATION PRACTICES: A BASIC HISTORY 3 (2012), https://bobgellman.com/rg-docs/rg-FIPshistory.pdf.

[30] ELECTRONIC PRIVACY INFORMATION CENTER, ALGORITHMIC TRANSPARENCY: END SECRET PROFILING, https://epic.org/algorithmic-transparency/; *see also* Margaret Hu, *Big Data Blacklisting*, 67 FLA. L. REV. 1735 (2015).

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

[32] *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

[33] FED. TRADE COMM'N, A CALL FOR TRANSPARENCY AND ACCOUNTABILITY, at i (2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

B. A Lack of Transparency

Objections to secrecy surrounding algorithms and machine learning sometimes have a very different meaning, often couched as a problem of algorithmic transparency. A transparency concern arises where the existence of a decision-making process is known, but the 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 such decisions are rendered. While this is perhaps the most frequent critique of algorithms and machine learning—that their inner-workings remain undisclosed or inaccessible[34]—this objection again 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 easily can—and quite often does—occur absent algorithms or machine learning.

There are sometimes valid reasons for companies to withhold details about the decision-making process. Where the discovery of some decision-making process holds financial and competitive value and where its discovery entailed significant investment or ingenuity, firms may claim protection for their discovery as a trade secret.[35] Trade secret protection only applies when firms purposefully restrict disclosure of proprietary methods,[36] thereby creating 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 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 the resistance to disclosure.[37]

[34] *See e.g.*, Robert Brauneis and Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J. L. & TECH. __ (forthcoming 2018); PASQUALE, *supra* note 14; Mikella Hurley & Julius Adebayo, *Credit Scoring in the Age of Big Data*, 18 YALE J.L. & TECH. 148, 196-98 (2016); Reddix-Smalls, *supra* note 12.

[35] Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. __ (forthcoming 2018); Brauneis & Goodman, *supra* note 34.

[36] Pasquale, *supra* note 12, at 237.

[37] *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

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.[38] 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.[39] 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 more or less easy to game than those developed using other methods—and this is not a question that can be answered in general.

C. A Lack of Expertise

Even requiring transparency will not be enough for accountability. A common version of the transparency demand is a call for disclosure of source code.[40] But 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."[41] Considerable problems remain even with direct access to the algorithms that drive decision-making because

secrecy in various technologies used in public infrastructure).

[38] Jane Bambauer & Tal Z. Zarsky, *The Algorithms Game* (draft on file with authors).

[39] Whether such manipulation is even possible will vary from case to case, depending on the degree to which the decision takes into account immutable characteristics and non-volitional 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 id.*

[40] 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; EPIC, *supra* note 30. Draft legislation in New York City also focused on this issue specifically, but the eventual bill convened a more general task force. *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.

[41] Ananny & Crawford, *supra* note 22, at 2.

in some cases the insights that Ananny and Crawford discuss are not present.[42] While source code disclosure might seem useful, the ability to make sense of the disclosed code depends on one's level of technical literacy; some minimal degree of training in computer programming is necessary to read code. (In reality, even that might not be enough.[43]) The problem, then, is greater than disclosures; in the absence of the required expertise to make sense of code, transparency may offer little of value to affected parties and regulators. Transparency into systems of decision-making is important, but incomplete.[44]

D. Inscrutability

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. 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 difficult or 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. The challenge in such circumstances is not a lack of awareness, transparency, or expertise, but the sheer scope and sophistication of the model.[45]

At first glance, complexity would seem to depend on the number of rules encoded by a model or the length of a rule (i.e. the number of

[42] Burrell, *supra* note 13.

[43] Kroll, et al., *supra* note 7, at 647; Desai & Kroll, *supra* note 40, at 5 ("[F]undamental limitations on the analysis of software meaningfully limit the interpretability of even full disclosures of software source code.")

[44] Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1254 (2008); Kroll, et al., *supra* note 7, at 639, 657-58.

[45] Burrell, *supra* note 13.

parameters that figure into the rule). But these properties can be specified more precisely. Four mathematical 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.[46] Linear models tend to be easier for humans to understand and interpret because the relationship between variables is stable, lending itself to straightforward extrapolation. In contrast, the behavior of nonlinear models can be far more difficult to predict, even when they involve simple mathematical operations like exponential growth.[47]

A monotonic relationship between variables is a relationship that is either always positive or always negative. That is, an increase in one variable consistently results in either an increase or decrease in the other. Monotonicity aids interpretability because it, too, permits extrapolation, guaranteeing that the value of a variable only moves in one direction. 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 and 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. 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.

The dimensionality of a model is the number of features it considers. Two-dimensional models are easy to understand because they can be visualized graphically with a standard plot (with the familiar *x* and *y* axes). Three-dimensional models also lend themselves to effective visualization (by adding a *z* axis). But we have no way to visualize models that have more than three dimensions. While humans can grasp relationships between multiple variables without the aid of a graph, people will struggle to understand the full set of relationships that the model has uncovered as the

[46] Mathematically, this means that the function is described by a constant slope, that it can be represented by a line. Yin Lou, Rich Caruana & Johannes Gehrke, *Intelligible Models for Classification and Regression.* PROC. 18TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 150 (2012).

[47] *See e.g.,* DEMI, ONE GRAIN OF RICE: A MATHEMATICAL FOLKTALE (1997).

number of dimensions grows. The more features that the model takes into account, the more difficult it will be to keep all the interactions between features in mind and thus predict how the model would behave given any particular input.

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 of three: simulating in one's mind how a model turns inputs into outputs.[48] As computer scientist Zachary Lipton explains, simulatability—the ability to practically execute the model in one's mind—is an important form of understanding a model.[49] Such simulations can be 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.

Simulation is a remarkably flat and functional definition of understanding, but it seems like a minimum requirement for any more elaborate definition of understanding.[50] But 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 sufficiently complex that humans are unable to perform this task, they have reached the point of inscrutability.

[48] Zachary C. Lipton, *The Mythos of Model Interpretability*, PROC. 2016 ICML WORKSHOP ON HUMAN INTERPRETABILITY IN MACHINE LEARNING 96, 98.

[49] *Id.*

[50] 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 13; 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 in photos). We return to one such example that detects wolves and huskies in Part II.B.2, *infra.* One thing the algorithm will usually learn to detect are 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 here. *See* Lipton, *supra* note 48, 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 hand-crafted by experts in the domain (e.g., length of employment), and accordingly will usually not face this problem.

E. Non-Intuitiveness

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 features to the decision at hand.[51] The problem in such cases is not a lack of transparency, technical expertise, or inscrutability, but an inability to weave a sensible story to account for the statistical relationships in the model.[52] While people might readily understand the statistical relationship that serves as the basis for decision-making, that relationship may defy intuitive expectations about the relevance of certain criteria to the decision at hand. 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.[53]

While it is clear that these specific statistical relationships serve as the basis for decision-making, why such statistical relationships should hold is mystifying. This is a crucial and consistent point of confusion; the demand for intuitive relationships is not the demand for transparency or accessible explanations. In social science, similar expectations are referred to as "face validity."[54] While such demands are not unique to algorithms and machine learning—there are many situations where one rightly expects coherence in human-made decisions—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.

With this in mind, critics have tended to pin this problem on the fact that machine learning operates on the basis of "mere correlation," which

[51] Deborah Gage, *Big Data Uncovers Some Weird Correlations*, WALL STREET J, (March 23, 2014), https://www.wsj.com/articles/big-data-helps-companies-find-some-surprising-correlations-1395168255; Quentin Hardy, *Bizarre Insights from Big Data*, N.Y. TIMES (March 28, 2012), https://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/

[52] *See* Brennan-Marquez, *supra* note 17, at 1280–97.

[53] Paul Ohm, *The Fourth Amendment in A World Without Privacy*, 81 MISS. L.J. 1309, 1317 (2012).

[54] *See generally* Ronald R. Holden, *Face Validity*, in CORSINI ENCYCLOPEDIA OF PSYCHOLOGY (2010).

frees it to uncover reliable, if incidental relationships in the data that might then serve as the basis for consequential decision-making.[55] While framed as an indictment of correlational analysis, it is really an objection to decision-making that rests on particular correlations that defy familiar causal stories[56]—even though these stories may be incorrect.[57] This has led to the mistaken belief that forcing decision-making to rest on causal mechanisms rather than mere correlations will ensure intuitive models.[58]

Causal relationships can be exceedingly complex and non-intuitive, especially when dealing with human behavior.[59] 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. If one considers all the different events that cause any one human decision: mood, amount of sleep, what the person ate that day, rational choice, and many other things that we cannot imagine, it becomes clear that causality is not particularly straightforward.[60] 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

[55] Kim, *supra* note 4, at 875; *see also* James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 173 (2017).

[56] *See* Brennan-Marquez, *supra* note 17, at 1280–97.

[57] *See* DANIEL KAHNEMAN, THINKING FAST AND SLOW 199-200 (2011) (discussing the "narrative fallacy"); 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.").

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

[59] *Id.*

[60] Attempts to model causation actually 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 401–428 (2d. ed. 2009).

reliability will not be a satisfying answer.

* * *

What the demand for intuitive relationships reflects is a desire to ensure that we have some way to assess whether the basis of decision-making is sound, both 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. Forcing a model to rely exclusively on features that bear a manifest relationship to the outcome of interest is a way to impose normative constraints on the model. On this account, well-justified decisions are those that rest on relationships that conform to familiar and permissible patterns.

This model 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 features with an intuitive relationship to the outcome.[61]

But intuitive relationships are not the only way to achieve the goal of an evaluable model. To the extent 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 goals. The remainder of this Article considers the different paths we might take to use explanations of machine learning models to regulate them. We start by describing the current approaches to solving inscrutability, which focus too narrowly on the machines themselves. An evaluation of the soundness of decision-making is an inherently human, subjective assessment, and cannot be resolved by describing the models alone.

[61] *See, e.g.*, Jiaming Zeng, Berk Ustun & Cynthia Rudin. *Interpretable Classification Models for Recidivism Prediction*, 180 J. ROYAL STAT. SOC'Y: SERIES A (STAT. IN SOC'Y) 689 (2017). Note that in this work and related work, the researchers limit themselves to features that are individually intuitively related to the outcome of interest. 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* Part III.B*, infra*.

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. And in Europe, concern about automated decisions has been part of data protection law for more than two decades, though the recently passed General Data Protection Regulation (GDPR)[62] has reinvigorated interest in those provisions. On the machine learning side, the sub-field of "interpretability"—within which researchers have been attempting to find ways to understand complex models—is over thirty years old.

What seems to emerge from the law and technical approaches 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 the technique. As a result, some legislation already exists that seeks to regulate by having systems explain themselves. In this Part, we discuss two examples of different legal systems and strategies that rely on different types of explanations. Credit scoring is an example of an inscrutable system that predates machine learning, and for which some regulation seeks explanations. Credit scoring is governed by two statutes: the Fair Credit Reporting Act (FCRA)[63] and the Equal Credit Opportunity Act (ECOA).[64]

[62] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. L 119/1, art. 13(f)(2), 14(g)(2), 15(1)(h)(requiring access to "meaningful information about the logic" of automated decisions) [hereinafter "General Data Protection Regulation" or "GDPR"].

[63] 15 U.S.C. § 1681, *et seq.*

[64] *Id.* § 1691, *et seq.*

Statistical credit scoring systems take information about consumers as inputs, give the inputs certain point values, add them to obtain a total score, and then make decisions based on that score. Each of these statutes require so-called "adverse action notices" that must include a statement of reasons for denials of credit or other outcomes based on credit.

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.[65] As the law has not yet taken effect, the import and proper interpretation of the requirement remain unclear. But in advance of the ultimate 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 other substantive rights under the GDPR and human rights law.[66]

1. FCRA, ECOA, and Regulation B

Before the 1950s, credit was not a large part of daily life in the United States. During the 1950s and 1960s, as credit became more popular, small credit bureaus arose, amassing large quantities of information about prospective credit applicants.[67] These bureaus would both "track peoples' names, addresses, and loan information" and "scour newspapers for notices of arrests, promotions, and marriages."[68] Once credit became more common, credit decisions were guided largely by the "three C's of credit": capacity, character, and collateral.[69]

[65] GDPR art. 13(f)(2), 14(g)(2), 15(1)(h)(requiring access to "meaningful information about the logic" of automated decisions).

[66] *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 116–120 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.

[67] ROBINSON + YU, KNOWING THE SCORE: NEW DATA, UNDERWRITING, AND MARKETING IN THE CONSUMER CREDIT MARKETPLACE 26 (2014), https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf.

[68] *Id.*

[69] Winnie F. Taylor, *Meeting the Equal Credit Opportunity Act's Specificity Requirement:*

By the late 1960s, the modern credit industry had begun to emerge, relying on amassed information and statistical models to predict creditworthiness.[70] While credit scoring was seen as a fairer, more objective way to make credit determinations,[71] consumers were nonetheless worried. Credit reports often contained incorrect or outdated information that credit reporting agencies (CRAs; e.g., Experian, Transunion, and Equifax) had no incentive to correct.[72] The industry was "secretive and enigmatic,"[73] and consumers had no idea who had access to their information or to what uses it was being put.[74]

Thus, in 1970, Congress passed FCRA[75] to begin to rein in the unregulated credit industry. FCRA was "the first information privacy legislation in the United States."[76] It limits to whom and for what credit reports can be disclosed,[77] allows consumers access to their credit reports,[78] and requires CRAs to employ procedures to ensure accuracy and dispute resolution.[79] FCRA was not initially concerned with how decisions were being made, but rather with the then-new phenomenon of amassing large quantities of information. Four years later, however, Congress passed ECOA,[80] and with it, took aim at the decision process. 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.[81]

Judgmental and Statistical Scoring Systems, 29 BUFF. L. REV. 73, 74 (1980).

[70] ROBINSON + YU, *supra* note 67, at 26.

[71] *See* Taylor, *supra* note 69, at 119.

[72] NATIONAL CONSUMER LAW CENTER, FAIR CREDIT REPORTING § 1.4.3 (8th ed. 2013) [hereinafter NCLC, FAIR CREDIT REPORTING].

[73] Lea Shepard, *Toward A Stronger Financial History Antidiscrimination Norm*, 53 B.C. L. REV. 1695, 1745 (2012).

[74] NCLC, FAIR CREDIT REPORTING, *supra* note 72, §§ 1.4.2-1.4.4.

[75] Pub. L. No. 91-508, tit. VI, 84 Stat. 1127 (1970) (codified as amended at 15 U.S.C. §§ 1681-1681x).

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

[77] *Id.* § 1681b.

[78] *Id.* § 1681g.

[79] *Id.* § 1681e(b)(2), § 1681i.

[80] Pub. L. No. 93-495, §§ 501-503 88 Stat. 1521 (1974) (codified at 15 U.S.C. § 1691(e))

[81] 15 U.S.C. § 1691.

ECOA introduced the adverse action notice requirement. When a creditor takes an adverse action against an applicant, the creditor must give a statement of "specific reasons" for the denial.[82] When FCRA later adopted the requirement, it expanded the notice to cover uses of credit information beyond credit decisions and beyond decisions made by creditors, including such decision-making as employment.[83] ECOA's notice requirement was implemented by the Federal Reserve Board in "Regulation B,"[84] which mandates that the "statement of reasons . . . must be specific and indicate the principal reason(s) for the adverse action." 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 creditors credit scoring system."[85] 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."[86] Though it is not necessary to use the form,

[82] *Id.* § 1691(d)(3); Taylor, *supra* note 69, at 82 ("For the first time, federal legislation afforded rejected credit applicants an automatic right to discover why adverse action was taken.")

[83] *Id.* § 1681m.

[84] 12 C.F.R. § 1002.1, *et seq.* ("Regulation B").

[85] *Id.* § 1002.9 (b)(2).

[86] 12 C.F.R. Pt. 1002, App. C ("Sample Form"). The 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

most creditors tend to report reasons contained on that form believing it to be a safe harbor.

Adverse action notices aim to serve the three purposes: 1) to alert a consumer that an adverse action has occurred; 2) to educate the consumer about how such a result could be changed in the future;[87] 3) to prevent discrimination.[88] As the rest of this part will show, these reasons are commonly cited reasons for targeting explanation as a means of regulation. 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.[89] But the relationship between adverse action notices and the other two rationales—consumer education and anti-discrimination—require further exploration.

Adverse action notices can sometimes be helpful for consumer education. As Winnie Taylor pointed out, writing 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.[90] Some explanations, such as "length of employment" are harder to understand or act on. Length of employment or home ownership are difficult to change. This suggests that an explanation of a specific decision may be informative, but it may not suggest an obvious path to an alternative

__Delinquent past or present credit obligations with others
__Collection action or judgment
__Garnishment or attachment
__Foreclosure or repossession
__Bankruptcy
__Number of recent inquiries on credit bureau report
__Value or type of collateral not sufficient
__Other, specify: ___

[87] *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.")

[88] S. REP. 94-589, 4, 1976 U.S.C.C.A.N. 403, 406 ("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.")

[89] *See supra* note 29 and accompanying text.

[90] Taylor, *supra* note 69, at 97.

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.[91] In a system like that, someone who comes up one point short could find himself with every factor listed as a "principal reason" for the denial. In one sense, this has to be correct because a positive change in any factor at all would change the outcome. But in another sense, choosing arbitrarily among equivalently valid reasons runs counter to the injunction 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).[92]

If the creditor tried to explain simply, it would leave information out, but if the creditor were to explain in complete detail how the system worked, 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.[93] 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."[94] 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

[91] *Id.* at 105-107.

[92] *Id.* at 123.

[93] *See* Ananny & Crawford, *supra* note 22, at 7 ("Transparency can intentionally occlude.")

[94] 12 C.F.R. Pt. 1002, Supp. I Para. 9(b)(2) ("Official Interpretations"). FCRA later codified the same limitation. 15 U.S.C. 1681g(f)(1)(C).

decision, especially when limited to four.[95] It is worth noting that modern credit systems appear not to be based on such complex models,[96] 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.[97] But the point is more general: approaches based on giving specific reasons for outcomes can fail where the system is too complex.

The notice fares worse as an anti-discrimination measure. By 1974, forcing hidden intentions into the open was a common technique for addressing discrimination. Just one year before ECOA's passage, *McDonnell Douglas Corp. v. Green* 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 non-discriminatory reason, and allows a plaintiff a chance to prove that the proffered reason is pretextual.[98] Just two years before that, the Supreme Court in *Griggs v. Duke Power Co.*[99] invented disparate impact doctrine, the purpose of which was arguably also to smoke out intentional discrimination where intent was hidden.[100] Thus, ECOA sought the same goal—to force decision-making into the open in order to prevent discrimination.

But while forcing stated reasons into the open captures the most

[95] 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, and "[i]f 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." *Id.*

[96] 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

[97] CAROL A. EVANS, KEEPING FINTECH FAIR: THINKING ABOUT FAIR LENDING AND UDAP RISKS 4–5 https://consumercomplianceoutlook.org/assets/2017/second-issue/ccoi22017.pdf?la=en; ROBINSON + YU, *supra* note 67, at 21.

[98] 411 U.S. 792, 805 (1973). 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.").

[99] Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

[100] Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 520-21 (2003) (discussing the "evidentiary dragnet" theory of disparate impact).

egregious forms of intentional discrimination, it does not capture much else. Although Regulation B bars collection of protected class information,[101] race, gender, and other features can be reliably inferred from sufficiently rich datasets.[102] Should creditors want to discriminate intentionally by considering membership in a protected class directly, they would have to affirmatively lie about such behavior lest they reveal obvious wrongdoing. Should creditors instead rely on known proxies for membership in a protected class, while they would have to lie about the true relevance of these features in predicting creditworthiness, they could honestly cite them as reasons for the adverse action. In neither case does the notice requirement place meaningful constraints on creditors, nor does it create additional or unique liability beyond those present in the anti-discrimination provisions of the rest of the regulation.[103]

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. Disparate impact doctrine 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 scoring system and no equally effective but less discriminatory alternative exists.[104] While ECOA does not expressly provide for a disparate impact theory of discrimination, case law suggests that it is very likely available.[105]

The adverse action notice approach has two specific shortcomings for a disparate impact case. First, the consumer only has access to her own specific outcome. She is told that she was denied because of one to four

[101] 12 C.F.R. § 1002.5

[102] Barocas & Selbst, *supra* note 4, at 692.

[103] John H. Matheson, *The Equal Credit Opportunity Act: A Functional Failure*, 21 HARV. J. LEG. 371, 388 (1984).

[104] 42 U.S.C. § 2000e-2(k)(1)(A). This description ignores the word "refuse," in the statute, but is probably the more common reading. Barocas & Selbst, *supra* note 4, at 709.

[105] The Supreme Court has not ruled that it is available, but most circuit courts that have considered it have permitted it. Hurley & Adebayo, *supra* note 34, at 193 (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 Housing & Cmty. Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507, 2518 (2015).

specific factors. Her single point of reference does not provide any understanding of the frequency of denials along protected class lines, so she cannot know whether there is a disparate impact. And with no 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.[106] 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 we discussed in Part I, one avenue to ensure that there is 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.[107] It is not the only way, as we will discuss in Part III, 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.

2. General Data Protection Regulation

In 2016, the European Union passed the GDPR, which takes effect May 25, 2018.[108] The GDPR is an EU-wide regulation that replaces the distributed system of data protection governed by the 1995 Data Protection Directive (Directive).[109] Both laws regulate automated decision-making,[110]

[106] Barocas & Selbst, *supra* note 4, at 702.

[107] *See* Part III.A.3, *infra*.

[108] GDPR, *supra* note 62, art. 99.

[109] Data Protection Directive, *supra* note 65.

[110] GDPR, *supra* note 62, 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 65, art. 15.

but in the 23 years of the Directive's existence, little jurisprudence has developed around that particular aspect of the law.[111]

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

> Article 22. Automated individual decision making, including profiling
>
> 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 [exceptions (a)-(c)].
>
> 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 subjects 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.
>
> 4. [omitted]

Articles 13-15 spell out a data subject's right to be informed about the data that data controllers have about them.[112] Articles 13 and 14 describe the obligations of data controllers to affirmatively notify data subjects about the uses of their information,[113] and Article 15 delineates the affirmative access rights that data subjects have to information about how their own data is used.[114] All three provide for the following information: "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."[115]

After the passage of the GDPR, scholars have begun to debate whether these requirements amount to a "right to explanation."[116] As one of

[111] Wachter, et al., *supra* note 21, at 19.
[112] *Id.* at 14.
[113] *See* GDPR, *supra* note 62, art.13-14.
[114] *See id.* art. 15.
[115] *Id.* art.13(2)(f), 14(2)(g), 15(1)(h).
[116] *See* Goodman & Flaxman, *supra* note 18; Wachter, et al., *supra* note 21; *See* Selbst &

us has argued elsewhere, that debate has been bogged down in proxy battles over what the phrase "right to explanation" means, but whether one calls it a right to explanation or not, requiring that data subjects have meaningful information about the logic has to mean something related to explanation.[117] Specifically, the GDPR's meaningful information requirement applies "to the data subject herself"[118] and "should be interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise his or her rights under the GDPR and human rights law."[119]

Importantly for this discussion, the regulation demands that the "meaningful information" must be about the *logic* of the decisions.[120] 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 overall purposes of the GDPR are much broader than FCRA and ECOA. The EU treats data protection as a fundamental right,[121] and Article 5 lists the following as principles the GDPR seeks to vindicate with respect to personal data: lawfulness, fairness and transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality, and accountability. Several of these principles are a restatement of the FIPs that have shaped privacy policy for decades.[122] But

Powles, *supra* note 66.; 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 (2017); Isak Mendoza & Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, *in* EU INTERNET LAW 77 (arguing that a right to explanation can be derived as a necessary precursor to the right to contest the decision); Sandra Wachter, Brent Mittlestadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR* __ HARV. J. L. & TECH __ (forthcoming 2018).

[117] *See* Selbst & Powles, *supra* note 66, at 233.

[118] *See id.* at 236.

[119] *Id.* at 242.

[120] GDPR, *supra* note 62, art.13(2)(f), 14(2)(g), 15(1)(h).

[121] *Id.* art. 1.

[122] Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward A Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 106-7 (2014). While different lists of FIPPs conflict, one prominent example is the OECD's list of eight: Collection Limitation Principle, Data Quality Principle, Purpose Specification Principle, Use Limitation

considered as a whole, including "lawfulness" and "fairness," 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 that govern all decision-making—that is, the entire model.[123] 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.

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

Data subject education becomes a lot more straightforward here, as a legal matter, if not technical. 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. Take the hypothetical system where a person missed on her credit application by one point, after the creditor totaled the point values from eight factors. While it might be impossible to point to four factors or fewer that were "principal reasons," the explanation of the logic—

Principle, Security Safeguards Principle, Openness Principle, Individual Participation Principle, and Accountability Principle. OECD, THE OECD PRIVACY FRAMEWORK 14-15 (2013), http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf

[123] The guidelines issued by 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 ("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.

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 particular rejected applicant.[124] Or 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. Of course, once machine learning enters the picture, and models become more complex, the limits on a technical ability to solve inscrutability may prevent these explanations from coming to fruition. But at least in theory, explanations of the logic are all that is needed for data subject education.

Turning to discrimination—which here is a stand-in for broader normative questions about model justification—logic-based explanations do a little better than outcome-based, but may not completely address the shortcomings. Any rule that is manifestly objectionable becomes visible, so that is an improvement over outcome-only explanations. And for rules that seem facially neutral, one might begin to speculate if they might 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.[125] There might not be anything that appears immediately objectionable, nor would it appear likely to generate a disparate impact, yet it still could. Or alternatively, a set of rules could appear objectionable or discriminatory, but ultimately be justified. It will often be impossible to tell without more, and the possibility of happening on a set of rules that lend themselves to intuitive normative assessment is 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 we demonstrated 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 further demonstrated that even meeting the requirements of the law does not automatically provide the types of explanations that would be necessary

[124] The Article 29 Working Party has, however, suggested that this approach is central to the "meaningful information" requirement. *See id.*

[125] *See* Part III.A.3, *infra.*

to assess whether decisions are well justified. And yet addressing the potential inscrutability of machine learning models remains a fundamental part of 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 what the field calls "interpretability."[126] Early research recognized and tried to grapple with the challenge of explaining the decisions of machine learning models such that people using these systems would feel comfortable acting upon them.[127] Practitioners and researchers have developed a wide variety of strategies and techniques to ensure that they learn 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. 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. This will certainly be the case when the phenomenon that machine learning tries to model is itself complex. This intuition suggests that favoring simplicity for the sake of interpretability will come at the cost of performance.[128]

While such views seem to be widely held,[129] over the past decade, methods have emerged that attempt to side-step these difficult choices altogether, promising to increase interpretability while retaining high performance.[130] The demand for explanations can be met with at least three

[126] Lipton, *supra* note 48; Riccardo Guidotti et al., *A Survey Of Methods For Explaining Black Box Models*, https://arxiv.org/abs/1802.01933 (forthcoming).

[127] Bruce G. Buchanan & Edward H. Shortliffe. *Rule-Based Expert System*, *in* THE MYCIN EXPERIMENT OF THE STANFORD HEURISTIC PROGRAMMING PROJECT __ (1984).

[128] Leo Breiman, *Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)* 16 STAT. SCI. 16, 199, 206-13 (2001).

[129] Henrik Brink & Joshua Bloom, *Overcoming the Barriers to Production-Ready Machine-Learning Workflows*, STRATA (2014); Such charts appear in government documents as well. DAVID GUNNING, EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI), DARPA-BAA-16-53 (Aug. 10, 2016), https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

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

different responses: 1) purposefully orchestrating the learning process such that the resulting model is interpretable; 2) applying special techniques after model creation to either approximate the model in a more readily intelligible form or identify features that are most salient for specific decisions; and 3) providing tools that allow people to interact with the model and get a sense for its operation.

1. Purposefully Building Interpretable Models

Where complexity might cause a model to become unwieldy, 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 features. 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. It is 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.[131] For example, decision tree algorithms 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 it would apply to all possible cases.

The experience of applying machine learning to real-world problems has led to widely held beliefs among practitioners about the relative interpretability of models that result from different learning methods. In particular, folk knowledge suggests that there is a trade-off between interpretability and accuracy.[132] Methods like linear regression generate models perceived as highly interpretable, but relatively low performing, while methods like deep learning result in high performing models that are exceedingly difficult to interpret. Though this is a commonly asserted and accepted trade-off, researchers have pointed out that such comparisons do

[131] David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 688-95 (2017).

[132] *See*, e.g., Breiman, *supra* note 128, at 208.

not rest on a rigorous definition of interpretability or empirical studies.[133] And yet such beliefs routinely guide practitioners' decisions when applying machine learning to different kinds of problems.[134]

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 at some point if they involve an exceedingly large number of branches and terminal leaves. Practitioners routinely set an upper bound on the number of terminal leaves to constrain the potential complexity of the model.[135] 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.[136] In so doing, they have been able to generate models that lend themselves to sensible explanation, but they have also knowingly forgone the additional accuracy that would come from a richer and more advanced analysis.[137]

Linear models remain the standard in industry because they allow companies to much more readily comply with the law. 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 play out different scenarios. 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 requested by the GDPR. At the same time, linear models also make the relative importance of different features immediately evident by assigning a specific numerical weight to each feature, which might allow 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 called regularization. Much like setting an upper limit on the number of branches in a decision tree, regularization methods allow model

[133] Alex A. Freitas, *Comprehensible Classification Models: A Position Paper.* 15 ACM SIGKDD EXPLORATIONS NEWSLETTER 1 (2014).

[134] *See* Lipton, *supra* note 48, at 4.

[135] *Id.*

[136] Hall, et al., *supra* note 96.

[137] *Id.*

complexity to be taken into account during the learning process by assigning a cost to excess complexity.[138] In doing so, model simplicity becomes an additional express objective alongside model performance—and the learning process can be set up in such a way as to find the optimal trade-off between these sometimes-competing objectives.[139]

Finally, the learning process can also be constrained in such a way that all features exhibit monotonicity. Monotonicity constraints are widespread in credit scoring because they make it easier for people to reason about how scores will change when the value of specific variables change and therefore allow creditors to automate the process of generating the reason codes required by FCRA and ECOA.[140] As a result of these legal requirements, creditors and others that use data-driven decision-making often have incentives to ensure that their models are interpretable by design.

2. Post Hoc Methods

An entirely different set of techniques for improved interpretability exist that do not place any constraints on the model-building process. Instead, these techniques begin with models learned with more complex

[138] One specific version of this method, widely used in practice, is called Lasso. Robert Tibshirani. *Regression Shrinkage and Selection Via the Lasso*, J. Royal Stat. Soc'y 267. Series B (Methodological) (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 patterns in the real world. Machine learning is only effective in practice when it successfully identifies robust patterns in the training data while also ignoring patterns that are just artifacts of the particular sample of cases assembled in the training data. Lasso increases accuracy by forcing the learning process to ignore relationships that are relatively weak, and therefore more likely to be artifacts of the specific examples that happened to be in the training data. Because Lasso works by strategically removing unnecessary features, in many real-world applications, the technique can simultaneously improve interpretability (by reducing complexity) and increase performance (by helping to avoid overfitting). Where it applies, this demonstrates that improved interpretability might not always come at the cost of performance. But where potential overfitting is not a danger, regularization methods will indeed result in degradations in performance.

[139] Gleicher, *supra* note 130.

[140] Hall, et al., *supra* note 96. Monotonicity allows creditors to rank order variables according to how much the value of each variable in an applicant's file differs from the value of such variables for the ideal customer—and the top four variables can function as reason codes.

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."[141] Once the learning process decides to introduce a particular branch, the method does not permit walking back up the branch. Therefore, if there is a relationship between items on two different branches, it will not be discovered. More complex learning methods, like support vector machines or neural networks, lack the same limitation, but they do not result in models as interpretable as decision trees. Nonetheless, rules that cannot be *learned* with simpler methods can often still be *represented* effectively by simpler models. Techniques like rule extraction can allow simple models to "cheat" because the answers that simpler learning method would otherwise miss are known ahead of time.

These methods are costly and do not have universal success. Practitioners must invest a considerable amount of time and effort to adapt and apply these techniques to their particular task. And despite practitioners' best efforts, replicating the performance of more complex models in a simple enough form might not be possible where the phenomena are themselves particularly complex. For example, approximating a model developed with deep learning in a decision tree might require too large a number of branches and leaves to be understandable in practice.[142]

When these methods work well, they ensure that the entire set of relationships learned by the model can be expressed concisely, without giving up much performance. Accordingly, they serve a similar role to the interpretability-driven design constraints discussed above. When they do not work as well, arriving at an interpretable model might necessitate sacrificing part 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 what would have been learned

[141] STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 92 (3d. ed. 2014).

[142] *See* Lipton, *supra* note 50, at 98.

with simple methods alone.[143] This helps to explain why such methods have been adopted in practice.

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. 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.[144]

These tools seem well suited for the task set by ECOA, FCRA, or future similar outcome-oriented approaches: explaining the principal reasons that account for the specific adverse decision. As we will discuss further 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: either determining the relative contribution of different features or identifying the features whose values would have to change the most to change the outcome.[145] 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

[143] Johan Huysmans, Bart Baesens & Jan Vanthienen, *Using Rule Extraction to Improve the Comprehensibility of Predictive Models* (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=961358

[144] David Baehrens, et al., *How to Explain Individual Classification Decisions*.11 J. MACHINE LEARNING RESEARCH 1803 (2010); Andreas Henelius, et al.. *A Peek Into the Black Box: Exploring Classifiers by Randomization*, 28 DATA MINING AND KNOWLEDGE DISCOVERY 1503 (2014); Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 KNOWLEDGE AND INFO. SYSTEMS 95 (2018); Marco Tulio Ribeiro, Sameer Singh, & Carlos Guestrin. *Why Should I Trust You?: Explaining the Predictions of Any Classifier. In* PROC. 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1135 (2016); Anupam Datta, Shayak Sen, & Yair Zick, *Algorithmic Transparency Via Quantitative Input Influence: Theory And Experiments With Learning Systems*, *in* SECURITY AND PRIVACY (SP), 2016 IEEE SYMPOSIUM 598 (2016).

[145] These methods are generally sensitive to interactions among variables and are able to measure indirect as well as direct influence. *See, e.g.*, Adler, et al., *supra* note 144; Datta, et al., *supra* note 144; JULIUS ADEBAYO, FAIRML: AUDITING BLACK-BOX PREDICTIVE MODELS (2017), http://blog.fastforwardlabs.com/2017/03/09/fairml-auditing-black-box-predictive-models.html.

make these relationships sensible to a human overall, they will provide a well-ordered list of reasons that explain a specific decision.

Unfortunately, however, 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 of them in an explanation for a particular decision is not likely to be terribly helpful. This is the machine-learning version of Taylor's hypothetical credit example. 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.[146] But the only alternative in these cases—arbitrarily listing fewer reasons than the correct number—is also unsatisfying when they are all 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 in 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, people will have a tendency to extrapolate from explanations of specific decisions to like cases, but the model—especially a complex one—may have a very different basis for identifying like cases. These methods offer explanations that apply only to the case at hand, and cannot be extrapolated to decisions based on other input data.

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 bear a strong resemblance to the more rigorous tools developed within computer science. This can take two quite different

[146] This might come as a surprise, given how well this approach works when applied to deep learning models, but recall that explanations in the case of object recognition take the form of visualizations: highlighting the specific pixels in an image that would have to change to change the classification.

forms. One is the type proposed by Danielle Citron and Frank Pasquale,[147] and implemented, for example, by Credit Karma.[148] 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, among others. A person using the interface can then select each of these to see how each action would affect his credit score. This does 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.

The second is more complicated and abstract. Mireille Hildebrandt has proposed something she terms "transparency-enhancing technologies."[149] Hildebrandt envisions an interface that would allow people to adjust the value of multiple features at the same time in a model, with the goal of providing people a loose sense of the relationship between these features and some outcome as well as the relationship between the features themselves. The goal of this type of technology is not to tell the user what changes in his results specifically, but allow him to get a feel from an arbitrary starting point.

While regulators have expressed interest in this idea,[150] it poses a difficult technical challenge. The statistical relationships at work in these models may be sufficiently complex that no consistent rule would become evident by tinkering with adjustable sliders, for example. Models might involve a very large number of inputs with complex and shifting interdependencies such that even the most systematic tinkering would quickly generate outcomes that would be difficult for a person to explain in a principled way.

Where models are simple enough, these approaches seem to get at the educational goals of both ECOA and the GDPR by allowing data

[147] *See* Citron and Pasquale, *supra* note 7, at 28-30 (discussing "interactive modeling").

[148] *See* CREDIT SCORE SIMULATOR, https://www.creditkarma.com/tools/credit-score-simulator.

[149] Mireille Hildebrandt, *A Vision of Ambient Law*, INFO. TECH. & SOC'Y COLLOQUIUM (2009). *See also* NICHOLAS DIAKOPOULOS, ALGORITHMIC ACCOUNTABILITY REPORTING: ON THE INVESTIGATION OF BLACK BOXES, http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf;

[150] INFORMATION COMMISSIONER'S OFFICE, BIG DATA, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA PROTECTION 87-88, https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

subjects to gain an intuitive feel for the system. But ironically, this would be accomplished by complying with neither law, because a person is unlikely to be able to give a specific reason for denial or an account of the logic after playing with it, even if they feel that they understand the system better afterward.

One danger of this approach 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 that they generated by playing with the system. But 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 (to give people a feel without disclosing the algorithm, for example), but for truly inscrutable systems, they could well be dangerous.

* * *

Remarkably, the techniques available within machine learning for ensuring interpretability correspond almost perfectly to the different types of explanation called for in both existing and forthcoming 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, either an expert (e.g., a regulator) or lay person (e.g., a person affected by the decision). Laws 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. But both techniques have their limitations. If highly sophisticated machine learning tools continue to be used, interpretability may simply be difficult to achieve in some instances, especially when the phenomena at issue are themselves complex. And 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 (e.g., when dealing with deep learning models).

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 or whether the decisions are justifiable.

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. This has certainly been the focus of legal and technical approaches, as we demonstrated in the prior Part. As far as we can tell, scholars, technologists, and policymakers have three different beliefs about the value of opening the black box. The first is a fundamental question of autonomy, dignity, and 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 by explaining the model, we can recognize its flaws.

But 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 the third relies heavily on the expected power of intuition. In this Part we demonstrate that an exclusive focus on the black box makes it appear that for those concerned with the justification for decision-making, the goal of explanation is to find a way to bring intuition to bear in deciding whether the model is well justified. We then explain both the power and limitations of an approach that relies on intuition.

A. The Value of Opening the Black Box

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.[151] There is a fundamental difference between wanting an explanation for its own sake, irrespective of what the specific explanation is, 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*,[152] a story about faceless bureaucracy that makes consequential decisions about people for which they have no input and no understanding.[153]

This concern certainly motivates some of the concern of lawmakers and scholars. In his article *Privacy and Power*, Dan Solove refers to this as a "dehumanizing" state of affairs described by the "powerlessness and vulnerability created by people's lack of any meaningful form of participation" in the decision.[154] David Luban, Alan Strudler, and David Wasserman argue that "a 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 individual's ignorance of what the fulfillment of his or her duty may entail."[155] The concerns of dignity and personhood certainly motivate the data protection regime in Europe,[156] if less directly the law in the United States.[157]

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

[151] 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").

[152] FRANZ KAFKA, THE TRIAL (1925).

[153] *See* Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1398 (2001) (arguing that Kafka's *The Trial* is a better metaphor than George Orwell's *1984* for modern anxieties over data).

[154] *Id.* at 1423.

[155] David Luban, Alan Strudler & David Wasserman, *Moral Responsibility in the Age of Bureaucracy*, 90 MICH. L. REV. 2348, 2355 (1992).

[156] Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUDIES SCI. 216, 223-24.

[157] *See* James Q. Whitman, *The Two Western Cultures of Privacy: Dignity Versus Liberty*, 113 YALE L.J. 1151 (2004).

personhood rationale can be converted to a more actionable legal issue, it is reflected in the concept of "procedural justice," 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.[158] 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 they find their procedural justice concerns satisfied than if they have their preferred outcome in the proceeding.[159] Procedural justice, Tyler argues, is necessary on a large scale because it allows people to buy in to the legal system, and comply with the law, both of which are essential parts of a working legal system.[160] Presumably, to the extent automated decisions can be legally or morally justified, people will have to accept them rather than have them imposed externally, and 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. It is difficult to quantify the inherent value of explanation or to compare it to other concerns. To the extent there are genuine tradeoffs between explanation and other normative values such as accuracy or fairness, we believe the inherent value of explanation does not automatically trump competing considerations. Nor does noting the inherent value provide much guidance as to the type of explanation required. While the inherent value cannot be ignored, it is not sufficient to end the discussion at this point.

2. Explanation as Enabling Action

Some scholars and policymakers have focused on explanation as a means to enable action in the consumer or data subject. This reflects the desire for consumer education that justified credit scoring regulations. While we do not intend to claim that consumer or data subject explanation is the sole focus of the scholars involved in this discourse, we believe the desire for actionable explanations is driving a good deal of this work.

[158] Tom R. Tyler, *What Is Procedural Justice?: Criteria used by Citizens to Assess the Fairness of Procedures*, 22 L. & SOC'Y REV. 103, 132.

[159] *See, e.g.*, Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 CRIME & JUSTICE 287, 291 (2003).; Tyler, *supra* note 158 at 128.

[160] TOM R. TYLER, WHY PEOPLE OBEY THE LAW (1990).

Within this subset of the total work, the scholarship has broken into two related debates. The first is whether the goal of black box explanation is outcome- or logic-driven, and the second is about how to best explain outcomes in an actionable way.

The divide between outcome- and logic-based explanations originates with an Article by Sandra Wachter, Brent Mittlestadt, and Luciano Floridi.[161] These scholars split explanations between "system functionality" and "specific decisions."[162] As they 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."[163] Aside from a few details, this framework roughly corresponds to our discussion of outcome- and logic-based explanations. In a responsive Article, one of us argued 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 be more useful.[164] This mirrors the debate in the technical community about the best way to understand the meaning of interpretability. As we described in Part II.B, the main split within the technical community is whether to aim for interpretable models or to account for specific decisions.

As the discussion has evolved in the legal scholarship, new work has seemingly converged on the belief that explaining specific outcomes is the right approach. The debate has therefore shifted to the different methods by which specific decisions can be explained, of which there are many. For example, a 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, though, we generally mean the reasons or justifications for that particular outcome, rather than a description of the decision-making process in

[161] Wachter, et al., *supra* note 21.

[162] *Id.* at 78. They actually insert explanations into a 2x2 grid: they can be either ex ante or ex post and can be either explanations of system functionality or specific decisions. Only the latter part is relevant to the discussion at hand.

[163] *Id.*

[164] Selbst & Powles, *supra* note 66, at 239; *see also* Citron & Pasquale, *supra* note 7, at 26 (focusing on the "the logics of predictive scoring systems").

general."[165] They propose three different ways to examine a specific decision: the main factors in a decision, the minimum change required to switch the outcome of a decision, and explanations for similar cases with divergent outcomes or divergent cases with similar outcomes.[166]

Wachter, Mittlestadt, and Chris Russell are narrower still, focusing on counterfactual explanations that represent "the smallest change to the world" that would result in a different answer.[167] 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.[168]

In the beginning of their article, Wachter, Mittlestadt, and Russell discuss three rationales for explanations: to help an individual understand the decision, to help contest the decision, and to enable action to create a better outcome.[169] These are similar to the three we mention here. But when they apply their suggested intervention of counterfactual explanations, it is clear that most of the value comes from the last rationale: actionable explanations. Their discussion of how counterfactuals aid understanding simply argues that as a matter of positive law, the GDPR requires almost nothing except a "meaningful overview," which can be encapsulated via pictorial "icons" about the type of processing done, and that because counterfactual explanations offer *something* specific to the data subject, they aid understanding more. If their interpretation of the law is correct,[170] then

[165] FINALE DOSHI-VELEZ, ET AL., ACCOUNTABILITY OF AI UNDER THE LAW: THE ROLE OF EXPLANATION, https://arxiv.org/abs/1711.01134 at 2.

[166] *Id.* at 3.

[167] Wachter, Mittlestadt & Russell, *supra* note 116.

[168] *Id.* at *12–16. Distance metrics are a way to solve this problem. Hall et al. describe another distance metric that is used in practice. Hall, Phan & Ambati, *supra* note 96. 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, for example, 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 is all about attempting to resolve this question.

[169] Wachter, Mittlestadt & Russell, *supra* note 116, at *5.

[170] The positive law debate about the right to explanation is not the subject of this Article, but suffice it to say, despite the certainty with which the authors state their

offering more than nothing is not saying much. Meanwhile, in the later discussion of using counterfactuals to contest the decision, the authors admit that in order to contest a decision, it is likely necessary to understand the logic of a system, rather than be given a counterfactual explanation.[171] The real value, then, of their intervention, is to better allow data subjects to alter their behavior, when the counterfactual in question suggests that the decision is made based on alterable characteristics.

Lilian Edwards and Michael Veale took a different approach, thinking about "model-centric" and "subject-centric" explanations.[172] They define these terms as follows: "Model-centric explanations (MCEs) provide broad information about a ML 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."[173] Thus, SCEs are another way to explain specific outcomes. They further differentiate SCEs as sensitivity-based, case-based, demographic-based, and performance-based, but do not elaborate after defining the terms.[174] Edwards and Veale do set themselves apart from the other scholars discussed here because their MCEs are in part explanations outside the black box, and bear some similarity to what we seek to accomplish in Part IV.

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 that 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. 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.[175]

interpretation of the law, it is just one interpretation. *See* sources cited *supra* note 116.

[171] *Id.* at *38. Their one example where a counterfactual can lead to the ability to contest is a similarly atypical case to those we discuss in Part III.C, *infra*.

[172] Edwards & Veale, *supra* note 116, at 55–59.

[173] *Id.* at 55–56.

[174] *Id* at 58.

[175] 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*

3. Explanation as Exposing a Basis for Evaluation

The final presumed value of explanation is that it will reveal some basis to question the validity of or normatively object to decision-making. As Pauline Kim has observed:

> 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.

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

The work of Rich Caruana et al. begins to answer that question.[176] They discovered that a model trained to predict complications from pneumonia had learned to associate asthma with a reduced risk of death.[177] 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.[178] What the model missed, however, was that these patients regularly monitored their breathing causing them to go to the hospital earlier, and once at the hospital, they were thought to be higher risk, so they received more immediate and focused treatment.[179] Caruana et al. 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 it has uncovered to determine if they correspond with domain knowledge. Thus, the purpose of explanation is to permit a

Daniel J. Solove, *Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013). In reality, this often fails 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).

[176] Rich Caruana et al., Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission, Proc. 21th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, 1721 (2015).

[177] *Id.* at 1721.

[178] *Id.*

[179] *Id.*

check against 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 perceived simply 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 obvious story we might be able to tell.

Another example of this view of explanation is the approach to interpretability known as Local Interpretable Model-Agnostic Explanations (LIME).[180] It has generated one of the canonical examples of the value of interpretability in machine learning. The authors investigated a model trained using deep learning, which was designed to tell wolves and huskies apart in photographs. Using LIME, they discovered that the model did not primarily rely on the animals' distinguishing features, but on whether snow appeared in the background of a photo.[181] 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, we can tell a story about why this mistake occurred: wolves are more likely to be found in an environment with snow on the ground. (Note: this story may not actually be true, but the important point is that we can convince ourselves it is.[182])

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

[180] Ribeiro, et al., *supra* note 144. This is one of the methods described in Part II.B.2, *supra*.

[181] *Id.* at 1142–43. This is a textbook example of overfitting the training data.

[182] In fact, at the time of writing, as we discussed the example and before consulting the original reference, we disagreed on whether the wolves or huskies were the ones pictured in snow. This goes to show that the story would have been equally compelling if the error had been reversed.

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 we offer come from work by James Grimmelmann and Daniel Westreich[183] and by Kim. Grimmelmann and Westreich imagine a scenario in which a model learns to distinguish between job applicants on a basis—musical taste—that is both correlated with job performance and membership in a protected class.[184] They further stipulate that job performance varies by class membership.[185] 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.[186]

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, which they assume to be membership in a protected class unless it can be shown otherwise.[187] The problem with this reasoning is that the model might not be learning protected class membership, but a latent variable that explains the relationship between musical taste and job performance. By assuming someone should be able to tell a story about such a variable, they—like 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. She cites a study stating that employees who installed new web browsers stay longer on their job.[188] She then also

[183] Grimmelmann & Westreich, *supra* note 55.

[184] *Id.* at 166–167.

[185] *Id.* at 167.

[186] 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.

[187] Grimmelmann & Westreich, *supra* note 55, at 173.

[188] Kim, *supra* note 4, at 922.

begins to speculate about the latent variable that would explain the relationship. (So too did the chief analytics officer in the company involved, in an interview.[189]) To Kim, what determines whether the relationship is "substantively meaningful" rather than a mere statistical coincidence is whether we can successfully tell such stories. Like Grimmelmann and Westreich, for Kim, if no such story can be told, and the model has a disparate impact, it should be illegal. 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 presumed values of explanation, the "conversation" approach does have the ultimate goal of evaluating whether the basis of decision-making is sound or 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.[190]

These two approaches are highly related and simply differ by how they treat strange cases by default. In the case of the two technical examples, the assumption is that obviously *flawed* relationships will present themselves and should be overruled; those for which there is no intuition 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 (they do allow for the possibility of defeating such a presumption). Both are forced rely on default rules to

[189] Joe Pinsker, *People Who Use Firefox or Chrome Are Better Employees,* THE ATLANTIC (Mar. 16 2015) https://www.theatlantic.com/business/archive/2015/03/people-who-use-firefox-or-chrome-are-better-employees/387781/ ("'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.'")

[190] 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, *see supra* note 58, so it may not satisfy the same normative desires that intuition seems 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. By definition, the reasons that models learned from data make decisions in the way they do is because they reflect the particular patterns in the data on which the models were trained. But this cannot be a sufficient explanation for why a decision is made the way it is if there are broader normative concerns at stake. Evaluating whether some basis for decision-making is fair, for example, will require tools that go beyond standard technical tests of validity already applied to the model during its development.[191] While the law gives these tests some credence, we have shown in prior work that reliance on accuracy is not normatively adequate with respect to machine learning.[192]

For many, the presumed solution is requiring machine learning models to be intelligible.[193] What the prior discussion demonstrates, though, is that this presumption works on a very specific line of reasoning, based on the idea that with enough explanation, we can bring intuition to bear in evaluating decision-making. As Kim observes, "[e]ven 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."[194] Her response

[191] 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. David J. Hand, *Classifier Technology and the Illusion of Progress*, 21 STAT. SCI. 1 (2006). In contrast, the law has broader concerns than real-world performance that demand very different justifications for the basis of decision-making encoded in machine learning models.

[192] 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.").

[193] Kim, *supra* note 4; Grimmelmann & Westreich 55, *supra* note 59; Brennan-Marquez, *supra* note 17.

[194] Kim, *supra* note 4, at 922.

is to constrain the model to features than bear an intuitive relationship to the outcome.[195]

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.[196] But there is a difference between a manifest relationship of a model to job performance and a manifest relationship or nexus between a particular *feature* and 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.[197] People who advocate for a nexus between features and the outcome are dissatisfied with a purely statistical test. This dissatisfaction manifests as an inability to normatively evaluate the model, even though it is statistically valid.

Regulatory guidance evinces similar reasoning. In 2011, the Federal Reserve issued formal guidance on model risk management.[198] The purpose of the document was to expand on prior guidance that was limited to model validation.[199] The guidance notes that models "may be used incorrectly or inappropriately" and that banks need diverse methods to evaluate them beyond statistical validation. Among other recommendations—which we will discuss in Part IV—the guidance recommends "outcomes analysis,"

[195] *Id.*; *Cf.* Nick Seaver, *Algorithms as Culture*, BIG DATA & SOC'Y, Jul.–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.")

[196] 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,' and/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))).

[197] Barocas & Selbst *supra* note 4, at 708.

[198] Federal Reserve Board, Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4 2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf.

[199] *Id*. at 2.

which calls for "expert judgment to check the intuition behind the outcomes and confirm that the results make sense."[200]

In an advisory bulletin about new financial technology, the Federal Reserve Board recommended that individual features have a "nexus" with creditworthiness in order to avoid discriminating in violation of fair lending laws.[201] In their view, a nexus enables a "careful analysis" about the features assigned significance in a model predicting creditworthiness. 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.[202] The business necessity defense asks whether the particular decision-making mechanism has a tight enough fit with the legitimate trait being predicted,[203] 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 Guidance is suggesting 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,[204] the Guidance treats an intuition requirement as a prophylactic. This reasoning seems to underlie the recommendations of Kim and Grimmelmann and Westreich as well.

What should be clear by now is that intuition is the typical bridge by which we go from explanation to normative assessment. And this can be a good thing. Intuition is powerful. It is a ready mechanism by which we can bring considerable knowledge to bear in evaluating machine learning models. Such models are myopic, having visibility into only the data upon

[200] *Id.* at 13–14.

[201] Evans, *supra* note 97, at 4.

[202] 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.

[203] *See, e.g.*, Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 1010 (1988) (business necessity is about careful tailoring).

[204] *See* Barocas & Selbst, *supra* note 4, at 709-710 (discussing the problems with the "fix the model" approach to alternative practice claims).

which they were trained.[205] 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 our experience. In fact, intuition is so powerful that we cannot keep ourselves from speculating about latent variables or causal mechanisms when confronted by unexplained phenomena.

Intuition can also take the form of domain expertise, which further strengthens our 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 has learned something meaningful. What appears strange on its face is given little credence or subject to greater scrutiny. Crucially, intuition allows us to generate competing explanations that account for the observed facts and to debate their plausibility.[206] Such a practice might seem ad hoc, but questioning face validity is a fundamental part of the social scientific process. Discoveries that run counter to expectations—that defy face validity—can give rise to further exploration and experimentation. This often takes the form of generating hypotheses about a latent variable or causal mechanism that might account for the initial finding—giving rise to an iterative process where new findings inform further hypotheses.

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. And it can equally lead us to endorse false discoveries because they align with our existing beliefs.[207] Intuition lets us generate "just so" stories that make good sense of the presented facts, but would make equally good sense of different or contrary facts. Such stories may feel coherent, but are in reality unreliable. In fact, the rich literature on cognitive biases—of which the so-called "narrative fallacy" is a part—is really an account of the dangers of intuition.[208] While intuition is helpful for assessing evidently good and bad results, it is less

[205] Selbst, *supra* note 10.

[206] Brennan-Marquez, *supra* note 17; *see also* Michael Pardo & Ronald J. Allen, *Juridical Proof and the Best Explanation*, 27 L. & PHIL. 223, 230 (2008).

[207] Raymond S. Nickerson, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, 2 REV. GENERAL PSYCHOLOGY 175 (1998).

[208] *See generally* KAHNEMAN, *supra* note 57.

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 us 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. Machine learning has the power to detect the subtle patterns and intricate dependencies that can better account for reality.

If our interest in interpretability is either the inherent value of explanation or actionable explanations, 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 way there. Ideally, solving inscrutability will restore our ability to bring intuition to bear in our normative assessments of decision-making. But sometimes our intuitions will fail us, even when we've been able to build interpretable models. In such cases, we should consider how else to justify models. We should think outside the black box. We should get back to the question: "why are these the rules?"

IV. DOCUMENTATION AS EXPLANATION

Stopping with the black box engages intuition by short-circuiting the question of why the rules are what they are. But what would it look like for regulation to actually seek answers to that question? The answers cannot come from the black box itself. In a sense, this division is implied by the very concept of a black box. There is a set of explanations internal to the operation of the box itself, and a set of explanations about the design of the system and how the system will be used, that by necessity are external. In order to get external explanations, we have to ask the humans.

Until recently, a common and often accepted answer for why the rules were the rules was that "the data says so." By now, though, it is well understood that data are human constructs[209] and that subjective decisions pervade the process of creating a model and deciding how to act on its recommendations.[210] What models learn will always be at least in part an artifact of the way its developers conceived of the problem at hand and the appropriate way to build a model to solve this problem.

In order to use those answers, we need to 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 really asking about the institutional and subjective process behind its development—the choices that account for the final decision-making process. The guidance discussed in Part III begins to get at this by recommending documentation, but it appears to be mostly about validation—how to do it well, thoroughly, on an ongoing basis, and in preparation for a future legal challenge.[211] 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, etc.[212] Careful validation is essential and it is non-trivial.[213] 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.

[209] Lisa Gitelman & Virginia Jackson, *Introduction*, *in* Raw Data is an Oxymoron 1 (Lisa Gitelman, ed., 2013); danah boyd and Kate Crawford, Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon, 15 INFO., COMM'N & SOC'Y 662, 666–68 (2012).

[210] Barocas & Selbst, *supra* note 4, at 673; *see also* Seaver, *supra*, note 195, at 5.

[211] Federal Reserve Board, *supra* note 198.

[212] *Id.* at 5–16; *see also* Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PENN. L. REV. ONLINE 189, 196 (2017); Edwards & Veale, *supra* note 116, at 55–56.

[213] Barocas & Selbst, *supra* note 4, at 680-692.

To illustrate how each of these components work, consider credit scoring. What are the values embedded in credit scoring models and what constraints do developers operate under? Lenders will attempt to achieve different objectives with credits scoring at the outset.[214] Credit scoring could aim to ensure that all credit is ultimately repaid, thus minimizing default. Lenders could use credit scoring to maximize profit. A lender could also seek to find ways to offer credit specifically to otherwise overlooked applicants, as many firms engaged alternative credit scoring seek to do. Each of these different goals reflect different core values. But other value judgements 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 enter 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.

Credit scorers will also face constraints and tradeoffs. For example, there might be limits on available talent with both domain expertise and the necessary technical skills to build models. Or models might be better informed if there were infinite data available, but there are practical challenges to collecting so much data. Ultimately, both tradeoffs are issues of cost,[215] 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 tradeoffs is a judgment call every firm will have to make.[216] One other cost-related tradeoff 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.[217] 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

[214] *See generally*, MARTHA ANN POON, WHAT LENDERS SEE—A HISTORY OF THE FAIR ISAAC SCORECARD (2013) (unpublished dissertation), http://search.proquest.com/docview/1520318884.

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

[216] Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 998 (1988) (plurality opinion) (consideration of costs and other burdens relevant to a discrimination case).

[217] Poon, *supra* note 214, at 93-134.

a result, competition operates as a constraint that ultimately changes the decision process. Even though competition operates across all firms, it is still useful to have documentation stating what work the constraint is doing in individual project design.

The values of and constraints faced by a firm will lead to certain choices about how to build and use models. As we discussed in prior work, the subjective choices a developer makes include choosing target variables, collecting training data, labeling examples, and choosing features.[218] 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.[219] Needless to say, the act of developing models is quite complex and involves many decisions by the developers.

In the credit example, the values discussed above may show up 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 the very thing the model seeks to optimize, 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 choose the same likelihood-of-default target variable, their values are likely to show up in the type of data they collect; they would seek alternative data sources 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—e.g. cost and competition. The traditional credit scorers tend to not make the extra effort or spend the extra money to obtain the data needed to make predictions about people on the margins of society.[220] There is also regulatory uncertainty regarding the permissibility of new types of credit data.[221] Therefore, their models reflected the fact that the developers are more sensitive to cost and regulatory penalty than

[218] Barocas & Selbst, *supra* note 4, at 677-692.

[219] Lehr & Ohm, *supra* note 131, 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).

[220] CFPB, *Request For Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process*, CFPB-2017-0005, at 6.

[221] *Id.* at 8, 30-35.

inclusion.

But 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?[222] 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 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%.[223] 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% and 60%, leaving the final decision to individual loan officers. Or a similar creditor might adopt a policy where applicants that the model is not able to score with great 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

[222] The distinction between models and ultimate decisions is what the GDPR aims to get at with Article 22's prohibition on "decision[s] based solely on automated processing." ARTICLE 29 WORKING PARTY, *supra* note 123, at 19-22.

[223] 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 seem to come from the criminal justice space. 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. When Chicago sought funding for the initiative, they premised it on the idea that they would provide increased social services to those 400 people, but in the end only ended up targeted them more for surveillance. DAVID ROBINSON & LOGAN KOEPKE, STUCK IN A PATTERN: EARLY EVIDENCE OF "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 for COMPAS, the now-infamous recidivism risk score. Rather than be used for further incarceration, it was originally designed to figure out who would need greater access to social services upon reentry. Julia Angwin, et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

particular uses of models are actually not about models at all, but about whether certain policies are acceptable independent of whether they use machine learning.[224] Questions about justifying a model are often just questions about policy in disguise.[225] 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 would otherwise. This will create more profit because the underserved loan candidates will be more willing to pay a higher rate, but it is pretty clearly predation; interest rates are not being used to offset risk, but to extract maximum profit from vulnerable people.[226] 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 choices made by the developers against society's broader normative priorities, as expressed in law and policy. That is the context in which models are developed and should be judged. In order to accomplish this evaluation, then, documentation about the decisions that lie behind and become part of models must exist and be made available for scrutiny. Now that we understand what that information looks like, we can begin to think about how to ensure that it is accessible.

B. Providing the Necessary Information

Assuming the documentation exists, there are two ways it can become open to scrutiny. One is that the documentation is made publicly available from the start and the other is that it becomes accessible to oversight upon some trigger. The former is essentially an algorithmic impact statement (AIS),[227] a proposed variant of the original impact statements

[224] VIRGINIA EUBANKS, AUTOMATING INEQUALITY __ (2018).

[225] *See*, *e.g.*, *id.*; Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 99-101 (2017); Sonia Katyal, *Algorithmic Civil Rights* (draft on file with authors); Margaret Hu *Algorithmic Jim Crow*, 86 Fordham L. Rev. 633 (2017); Sandra G. Mayson, *Dangerous Defendants*, 127 YALE L.J. 490, 507–518 (2018)

[226] 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*, HOUSING POL'Y DEBATE (2013).

[227] Selbst, *supra* note 4, at 169–93.

required by the National Environmental Policy Act.[228] For the latter, the most common trigger is a lawsuit, in which documents can be obtained and scrutinized, and witnesses can be deposed or examined on the stand. In both methods, the coupling of existing documentation and 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 connection to machine learning, but their methodologies can be easily applied here.

An impact statement is a document designed to explain the process of decision-making and the anticipated effects of that decision, and to do so 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.[229] Impact statements are an idea that originated in 1970 with the National Environment Policy Act,[230] but have since been emulated repeatedly at all levels of government, in many substantive areas of policy.[231] For example, aside from environmental law, the federal government requires privacy impact assessments "when developing or procuring information technology systems that include personally identifiable information."[232] Individual states not only have their own legislation requiring environmental impact statements,[233] but also, racial impact statements for sentencing policy, for example.[234] Recently, led by the ACLU's "Community Control Over Police Surveillance" (CCOPS) initiative,[235] counties and cities have begun

[228] See 42 U.S.C. § 4332(C) (2012).

[229] Selbst, *supra* note 4, at 169

[230] 42 U.S.C. § 4321, *et seq*. (2012).

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

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

[233] *E.g.*, California Environmental Quality Act (CEQA), CAL. PUB. RES. CODE § 21000 *et seq*.

[234] Jessica Erickson, *Racial Impact Statements: Considering the Consequences of Racial Disproportionalities in the Criminal Justice System*, 89 WASH. L. REV. 1425, 1445 (2014); London, *supra* note **Error! Bookmark not defined.**, at 226–31.

[235] AN ACT TO PROMOTE TRANSPARENCY AND PROTECT CIVIL RIGHTS AND CIVIL LIBERTIES WITH RESPECT TO SURVEILLANCE TECHNOLOGY §2(B) https://www.aclu.org/files/communitycontrol/ACLU-Local-Surveillance-Technology-

requiring impact statements that apply to police purchases of new technology.[236]

One of us has argued that a future algorithmic impact statement (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 variable, 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.[237] 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 possibility of no action, and the effects of any potential mitigation procedures.[238]

In the typical American example of impact statements, they are public documents. Thus, a law requiring them would also require that the developers publish the document, and allowing for comments between the draft and final impact statements.[239] 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 extremely reluctant to publish an AIS that reveals operating strategy, perceived constraints, and even embedded values. Thus, it is also useful to consider a documentation requirement that allows the documentation to remain private, but available as needed.

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."[240] As Edwards and Veale discuss, the DPIA requirement is very likely to apply to machine learning,[241] and the assessments require

Model-City-Council-Bill-January-2017.pdf (ACLU CCOPS Model Bill).

[236] SANTA CLARA COUNTY, CAL. CODE DIV. A40 (2016).

[237] Selbst, *supra* note 4, at 172–75.

[238] *Id.*

[239] *Id.* at 177.

[240] GDPR, *supra* note 62, art. 35.

[241] Edwards & Veale, *supra* note 116, at 77–78.

"appropriate technical and organizational measures" to protect data subject rights.[242] In Europe DPIAs are private documents, for which nothing but a summary need be made public.[243] 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.[244]

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 rules that enable parties to litigation to question human witnesses and examine documents they have created. These are all chances to directly ask the designers of decision systems 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—differs from the use of intuition as a method of justification.

Imagine a new alternative credit scoring system that relies on social media data.[245] This model assigns significance to data points that are unintuitive, but that reliably predict default. Suppose the model also evinces a disparate impact along racial lines, which was revealed by investigative journalists.

Black applicants denied credit then bring suit under the substantive non-discrimination provisions of ECOA. Assuming, reasonably, that the judge agrees that disparate impact is a viable theory under ECOA,[246] the case will turn on the business necessity defense. Thus, in order to figure out if 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.

[242] GDPR, *supra* note 62, art. 35.

[243] 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, Art. 29, WP 248, at 18 (Apr. 4, 2017).

[244] Edwards & Veale, *supra* note 116, at 78.

[245] *See, e.g.*, Astra Taylor and Jathan Sadowski, *How Companies Turn Your Facebook Activity Into a Credit Score*, THE NATION (May 27, 2015), https://www.thenation.com/article/how-companies-turn-your-facebook-activity-credit-score/

[246] *See* CFPB Bulletin 2012-04 (Fair Lending), at 2 (Apr. 18, 2012), https://files.consumerfinance.gov/f/201404_cfpb_bulletin_lending_discrimination.pdf.

Under an intuition-driven regime, such as that proposed by either Kim or Grimmelmann or 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.[247] Under their proposed regime, the defendant has not satisfied their burden, and she would be held liable.[248]

Under a regime of mandated documentation, however, other explanations could be used in the defense of the model. The engineer would be permitted to answer, not just that she cannot intuitively link the social media data to the creditworthiness, but why the model relies on the data in the first place. The documentation might show (or the engineer might testify), for example, that her team tested the model with and without the social media data, finding that using the data reduced the disproportionate impact of the model. (In fact, a recent Request for Information by the Consumer Financial Protection Bureau seems to anticipate such a claim.[249]) Alternatively, the documentation might demonstrate that the team considered other, more intuitive features that might be necessary for a more accurate and fairer model, but then discovered that such features were exceedingly difficult or costly to measure. The company then used social media data because it increased accuracy and fairness under the practical constraints faced by the company.

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

[247] 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.

[248] Grimmelmann & Westreich, *supra* note 55, at 170.

[249] CFPB , *supra* note 220, at 7–8.

difficult normative line,[250] and would likely require a case-by-case analysis. Intuition-based evaluation—and its reliance on default rules—would forbid the consideration of either of these motivations for using social media data, but both rationales should at least enter into the discussion.[251]

Whether accomplished through public documentation or private documentation with access, having to account for all the decisions made in the process of project inception and model development should reveal a number of subjective judgments than 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.[252] 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.

Some models will both defy intuition and resist normative clarity even with documentation. But documentation leaves open the possibility that we might develop other ways of asking whether this was a well-executed—intuitions about what constitutes best practice. For example, 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.[253] Documentation of the decisions taken

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

[251] 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). Normally, the response of "fix the model" 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. But with documentation, the alternatives will be plain as day, because that is exactly what has been documented.

[252] Barocas & Selbst, *supra* note 4, at 715.

[253] William Smart, Cindy Grimm, and Woody Hartzog, *An Education Theory of Fault for*

will be also be necessary to such a regime.

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."[254] 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 of 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. But those that believe solving inscrutability provides a path to normative evaluation also fall short of the goal because they fail to recognize the role of intuition.

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

Just as restricting our 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

Autonomous Systems, (draft on file with authors)

[254] KAHNEMAN, *supra* note 57, at 185.

be a role for intuition, we will not always be able to use intuition to bypass the question of why the rules are the rules. Sometimes 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 our way of making decisions, we tend to focus too much on the technology itself, when we should be focused on the policy changes that either led to the adoption of the technology or were wrought by the adoption.[255] Quite aside from correcting one failure mode of intuition, then, the 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.

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 judgements about competing values under their real-world constraints. Only humans know the answer.

[255] *See generally* EUBANKS, *supra* note 224.