B. Interpretability in Machine Learning

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

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

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

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

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

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

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

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

1. Purposefully Building Interpretable Models

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

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

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

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

160. *See generally id.*

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

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

163. *See, e.g.*, *id.* at 81–82.

164. *See, e.g.*, *id.* at 82–83.

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

166. *See id.* at 81.

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

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

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

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

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

---

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

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

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

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

179. *Id.*

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

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

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

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

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

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

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

2. Post Hoc Methods

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

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

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

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

---

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

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

187. *Id.*

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

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

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

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

192. *Id.*

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

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

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

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

---

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

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

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

197. *See generally* Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 KNOWLEDGE & INFO. SYSTEMS 95 (2018); David Baehrens et al., *How to Explain Individual Classification Decisions*, 11 J. MACHINE LEARNING RES. 1803 (2010); Anupam Datta et al., *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*, *in* PROCEEDINGS OF THE 2016 IEEE SYMPOSIUM ON SECURITY & PRIVACY 598 (2016); Andreas Henelius et al., *A Peek into the Black Box: Exploring Classifiers by Randomization*, 28 DATA MINING & KNOWLEDGE DISCOVERY 1503 (2014); Marco Tulio Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, *in* PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1135 (2016).

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

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

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

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

3. Interactive Approaches

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

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

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

202. *See id.*

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

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

205. *Id.*

206. *Id.*

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

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

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

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

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

---

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

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

209. *See id.*

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

* * *

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

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

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

III. FROM EXPLANATION TO INTUITION

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

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

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

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

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

A. The Value of Opening the Black Box

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

1. Explanation as Inherent Good

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

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

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

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

214. FRANZ KAFKA, DER PROCESS (1925).

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

216. *Id.* at 1423.

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

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

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

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

---

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

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

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

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

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

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

2. Explanation as Enabling Action

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

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

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

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

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

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

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

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

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

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

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

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

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

229. *Id.* at 3.

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

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

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

233. *Id.* at 865.

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

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

3. Explanation as Exposing a Basis for Evaluation

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

---

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

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

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

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

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

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

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

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

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

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

239. Kim, *supra* note 4, at 922–23.

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

241. *Id.*

242. *Id.*

243. *Id.*

244. *Id.*

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

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

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

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

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

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

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

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

249. Kim, *supra* note 4.

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

251. *Id.* at 167.

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

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

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

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

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

---

254. *Id.* at 173.

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

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

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

258. *Id.*

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

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

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

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

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

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