The Algorithmic Explainability "Bait and Switch"

Boris Babic and I. Glenn Cohen

Forthcoming in the Minnesota Law Review

ABSTRACT

Explainability in artificial intelligence and machine learning ("AI/ML") is emerging as a leading area of academic research and a topic of significant regulatory concern. Indeed, a near-consensus exists in favor of explainable AI/ML among academics, governments, and civil society groups. In this project, we challenge this prevailing trend. We argue that for explainability to be a moral requirement – and even more so for it to be a legal requirement – it should satisfy certain desiderata which it currently does not, and possibly cannot. In particular, we will argue that the currently prevailing approaches to explainable AI/ML are (1) incapable of guiding our action and planning, (2) incapable of making transparent the actual reasons underlying an automated decision, and (3) incapable of underwriting normative (moral/legal) judgments, such as blame and resentment. This stems from the post hoc nature of the explanations offered by prevailing explainability algorithms. We will explain that these algorithms are "insincere-by-design," so to speak. And this renders them of very little value to legislators or policymakers who are interested in (the laudable goal of) transparency in automated decision making. There is, however, an alternative – interpretable AI/ML – which we will distinguish from explainable AI/ML. Interpretable AI/ML can be useful where it is appropriate, but represents real tradeoffs and in some instances (in medicine and elsewhere) adopting an interpretable AI/ML may mean adopting a less accurate AI/ML. We argue that it is better to face those trade-offs head on, rather than embrace the fool's gold of explainable AI/ML.

Authors.
Boris Babic, JD, PhD, is Assistant Professor of Statistical Sciences and Philosophy at the University of Toronto. He is also a Faculty Fellow of the Schwartz Reisman Institute for Technology and Society. His work was supported by a Social Sciences and Humanities Research Council of Canada Insight Grant (no. 435-2022-0325).

I. Glenn Cohen, JD, is a Deputy Dean and the James A. Attwood and Leslie Williams Professor of Law at Harvard Law School. He is also the Faculty Director, Petrie-Flom Center for Health Law Policy, Biotechnology & Bioethics. His work was supported by a grant from the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program supported by a Novo Nordisk Foundation grant (NNF17SA0027784). He serves on the bioethics advisory board of Illumina and on the bioethics council of Bayer.

Table of Contents

INTRODUCTION

From cars, to cardiology, to Chat GPT-4 our world is increasingly being shaped by Artificial Intelligence (AI) and even more specifically the sub-type of AI known as Machine Learning (ML). As algorithmic decision making systems relying on AI/ML models become more prominent across the legal, commercial and medical landscape,[1] there is an increasingly vocal push by policymakers to require that these algorithms be more explainable.[2] For example, many scholars argue that the EU General Data Protection Regulation (2016/679)[3] contains a "right to explanation" for algorithmically generated decisions.[4] Likewise, a major piece of Canadian legislation known as Bill C-27,[5] or the Digital Charter Implementation Act, 2022, would require organizations using an "automated decision making system" to provide an explanation of its prediction when requested by a significantly impacted individual.[6] Notably for our project, the Canadian bill explicitly states that an explanation must include "the *reasons or principal factors* that led to the prediction."[7] We will return to why reason giving is such an important idea below. Turning to the United States, while it has tracked behind the EU and Canada in digital regulation, the White House Office of Science and Technology Policy's "Blueprint for an AI Bill of Rights,"[8] released on October 4,

[1] Andrew Guthrie Ferguson, *Illuminating Black Data Policing*, 15 OHIO ST. J. CRIM. L. 503, 504–09 (2018) (policing); Arthur Rizer & Caleb Watney, *Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just*, 23 TEX. REV. L. & POL. 181, 195 (2018) (pretrial detention); Cary Coglianese & Lavi M. Ben Dor, *AI in Adjudication and Administration*, 86 Brook. L. Rev. 791, 802–04 (2021) (sentencing and parole); Sofia Ranchordás, *Empathy in the Digital Administrative State*, 71 DUKE L.J. 1341, 1359–60 (2022) (administration); Ashley S. Deeks, *Predicting Enemies*, 104 VA. L. REV. 1529, 1547 (2018) (warfare); William Magnuson, *Artificial Financial Intelligence*, 10 HARV. BUS. L. REV. 337, 349–50 (2020) (credit ratings); *id.* at 350–51 (fraud detection); *id.* at 351 (investment); Ifeoma Ajunwa, *An Auditing Imperative for Automated Hiring Systems*, 34 HARV. J.L. & TECH. 621, 623 (2021) (hiring); Dana Remus & Frank Levy, *Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law*, 30 GEO. J. LEGAL ETHICS 501, 512–29 (2017) (legal practice); George Maliha, Sara Gerke, I. Glenn Cohen & Ravi B. Parikh, *Artificial Intelligence and Liability in Medicine*, 99 MILBANK Q. 629, 629–30 (2021) (medicine).

[2] *E.g.*, Algorithmic Accountability Act of 2022, H.R. 6580, 117th Cong. (2022); *see* Edmund L. Andrews, *Congress Gets Serious About Artificial Intelligence*, STAN. UNIV. HUMAN-CENTERED A.I. (Mar. 8, 2021), https://hai.stanford.edu/news/congress-gets-serious-about-artificial-intelligence.

[3] Council Regulation 2016/679, 2016 O.J. (L 119) 1.

[4] Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision-making and a "Right to Explanation,"* ARXIV (Aug. 31, 2016), https://arxiv.org/abs/1606.08813; Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIV. L. 233, 234 (2017). However, not all scholars agree that the GDPR entails a requirement that decision making algorithms be explainable. *See* Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIV. L. 76 (2017); Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189 (2019); Sara Gerke et al., *Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare*, *in* ARTIFICIAL INTELLIGENCE IN HEALTHCARE 295, 301 (Adam Bohr & Kaveh Memarzadeh eds., 2020).

[5] Bill C-27, 44th Parliament, 1st Sess. (Canada 2022), https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading; Jennifer R. Davidson et al., *Bill C-27, Proposed Amendments to Canada's Federal Privacy Legislation Affecting Private Sector Organizations*, 35 INTELL. PROP. J. 71 (2022).

[6] Bill C-27, *supra* note 5, §§ 63(3), 63(4).

[7] *Id.* § 63(4) (emphasis added).

[8] Alondra Nelson et al., *Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age*, WHITE HOUSE OFF. SCI. & TECH. POL'Y (Oct. 4, 2022),

2022, mirrors the proposed Canadian digital charter of rights and freedoms by including, among other things, an anticipated requirement for "notice and explanation" of algorithmic decisions. While it is unclear what the notice and explanation requirement would entail, the authors of the blueprint state that the purpose of the notice and explanation is to allow one to understand "how and why [an automated system] contributes to outcomes that impact you."[9] It seems plausible that providing the requisite notice and explanation would require the proprietors of automated decision making systems to produce reasons or factors for why certain decisions were made the way they were, and that these factors should be transparent enough to allow for a review of the decision.[10]

This trend has not gone unnoticed by academics, many of whom champion the importance of transparency. For example, some authors argue that algorithmic decision making "gives rise to a right to explanation."[11] Sandra Wachter, while maintaining that the GDPR does not in general give rise to a right to explanation, believes the law should incorporate indirect ways of providing explanations without "opening the black box."[12] Indeed, "a near-consensus is emerging in favor of explainable AI/ML among academics, governments, and civil society groups."[13] There is now a very large literature on explainable

https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/.

[9] *Id.*

[10] In US administrative law, for example, while reviewing courts often give wide deference to agency decisions, *see* Baltimore Gas & Elec. Co. v. Nat. Res. Def. Council, Inc., 462 U.S. 87, 103 (1983); *cf.* Chevron, U.S.A., Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837, 844 (1984), the reviewing court must still be able to understand an agency decision well enough to determine whether it was based on a consideration of the relevant factors and reasons, *see* Sec. & Exch. Comm'n v. Chenery Corp., 318 U.S. 80, 94–95 (1943).

[11] Tae Wan Kim & Bryan R. Routledge, *Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach*, 32 BUS. ETHICS Q. 75, 75 (2022).

[12] Sandra Wachter et al., *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841 (2018); *see also* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 Colum. L. Rev. 1829, 1835–37 (2019), Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 Colum. L. Rev. 1851, 1863–64 (2019). *But see* Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NATURE MACH. INTEL. 206 (2019).

[13] Boris Babic, Sara Gerke, Theodoros Evgeniou & I. Glenn Cohen, *Beware Explanations from AI in Health Care*, 373 SCIENCE, art. no. abg1834, 2021; for examples of techniques used to render AI explainable, see Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, Proc. 22nd ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING (Aug. 2016), at 1135, https://doi.org/10.1145/2939672.2939778; Scott Lundberg & Su-In Lee, *A Unified Approach to Interpreting Model Predictions*, PROC. 31ST INT'L CONF. ON NEURAL INFO. PROCESSING SYS. (Dec. 2017), https://dl.acm.org/doi/10.5555/3295222.3295230; Cynthia Rudin & Joanna Radin, *Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition*, 1 HARV. DATA SCI. REV., art. no. 2, 2019, https://doi.org/10.1162/99608f92.5a8a3a3d.

AI/ML.[14] It claims that explainable AI/ML systems are more trustworthy,[15] easier to understand,[16] safer,[17] and more accountable/transparent.[18] For example, Scott Lundberg and Su-In Lee write: "[t]he ability to correctly interpret a prediction model's output . . . engenders appropriate user trust . . . and supports understanding of the process being modeled."[19] Similarly, Marco Tulio Ribeiro et al. write: "Understanding the reasons behind predictions . . . is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model."[20]

We do not disagree with these sentiments insofar as they suggest that these are desirable features of an AI/ML system. The problem, we will argue, is that explainable AI/ML models fundamentally fail to achieve these goals: these models fail to assist users in either correctly interpreting a model, or in understanding the true reasons or principal factors behind the model's predictions. Specifically, we argue that there are three cardinal shortcomings of current explainable AI/ML systems, stemming from three features of the explanations these models generate: they are not unique, they are not sincere, and they are produced after the fact. The three shortcomings are: First, explanations produced by explainable AI/ML algorithms purport to be action guiding, but they are not. We will explain why they are inadequate for guiding our behavior, or assisting us in planning about the future. Second, and related, explainable AI/ML algorithms purport to shine a light on the *actual* (or otherwise

---

[14] Deeks, *supra* note 12; Ribeiro et al., *supra* note 13; Lundberg & Lee, *supra* note 13; Babic et al., *supra* note 13; PROC. OF ICML 2021 WORKSHOP ON THEORETIC FOUNDATION, CRITICISM, AND APPLICATION TREND OF EXPLAINABLE AI (Jul. 26, 2021), https://arxiv.org/abs/2107.08821; Muhammad Suffian et al., *FCE: Feedback Based Counterfactual Explanations for Explainable AI*, 10 IEEE ACCESS 72363 (2022); Sindhu Ghanta et al., *Interpretability and Reproducability in Production Machine Learning Applications*, 2018 17TH IEEE INT'L CONF. MACH. LEARNING & APPLICATIONS 658; Andreas Holzinger et al., *Causability and Explainability of Artificial Intelligence in Medicine*, 9 WIREs DATA MINING & KNOWLEDGE DISCOVERY, no. e1312, 2019; Katie Atkinson et al., *Explanation in AI and Law: Past, Present and Future*, 289 ARTIFICIAL INTELL. 103387 (2020); MARCO IANSITI & KARIM R. LAKHANI, COMPETING IN THE AGE OF AI: STRATEGY AND LEADERSHIP WHEN ALGORITHMS AND NETWORKS RUN THE WORLD (2020); Erwan Le Merrer & Gilles Tredan, *Remote Explainability Faces the Bouncer Problem*, 2 NATURE MACH. INTELL. 529 (2020); *see also infra* notes 15–18.

[15] Ribeiro et al., *supra* note 13, at 1135–36; MARK COECKELBERGH, AI ETHICS 118–19 (2020); Mary-Anne Williams, *Explainable Artificial Intelligence*, *in* RESEARCH HANDBOOK ON BIG DATA LAW 318, 325–27 (Roland Vogl ed., 2021); Will Knight, *The Dark Secret at the Heart of AI*, 120 MIT TECH. REV., no. 3, May/June 2017, at 55, 61.

[16] Matt Turek, *Explainable Artificial Intelligence (XAI)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/explainable-artificial-intelligence; COECKELBERGH, *supra* note 15, at 120; *see also* Williams, *supra* note 15, at 327 (explaining differences between explainability and interpretability).

[17] Yan Jia, John McDermid, Tom Lawton & Ibrahim Habli, *The Role of Explainability in Assuring Safety of Machine Learning in Healthcare*, 10 IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING 1746 (2022); Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez & Matthieu Cord, *Explainability of Deep Vision-Based Autonomous Driving Systems: Review and Challenges*, 130 INT'L J. COMP. VISION 2424 (2022).

[18] *See generally* Finale Doshi-Velez et al., *Accountability of AI Under the Law: The Role of Explanation*, ARXIV (Dec. 20, 2019), https://arxiv.org/abs/1711.01134; *see also* Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans in the Loop*, 76 VAND. L. REV. (forthcoming 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4066781 (manuscript at 25-26) ("There is a growing body of caselaw where algorithmic decisions were invalidated on procedural due process grounds."); COECKELBERGH, *supra* note 15, at 122–23.

[19] Lundberg & Lee, *supra* note 13, at 1.

[20] Ribeiro et al., *supra* note 13, at 1135.

put, motivating) reasons behind a decision. If successful, this would help in garnering trust, encouraging usage, and better enable the review of decisions. But, we will argue, these models fail to identify the actual reasons for a decision, providing instead the "fool's gold" of a post-hoc explanation that may not underlie the actual decision. Finally, explanations can be valuable insofar as they can underwrite normative judgments – such as blame and praise – or what are sometimes called in philosophy the second person or Strawsonian reactive attitudes.[21] These attitudes are intimately related to evaluating an agent's quality of will – are they blameworthy or praiseworthy in their behavior? In assigning blame and praise for a human agent we usually need to know the reasons that motivate their behavior – why they did what they did.[22] But the kind of explanations that explainable AI/ML generates cannot help us to do this – they cannot help us understand *why* the automated decision was made the way it was, nor can they help us understand the actual reasons or factors that led to it. As a result, the explanations cannot let us know whether we are right to feel one of these reactive attitudes towards the algorithm.

While one contribution of this paper is to explain these three shortcomings of explainable AI, a second contribution is to show that there is an alternative that *does* satisfy the desiderata that supporters of explainable AI/ML argue for – what we and others call "interpretable" AI/ML.[23] We will explain what it is, and how it differs from explainable AI/ML. In short, interpretable AI/ML uses simple, and usually additive, models which are intuitive and transparent, such as linear regressions and shallow decision trees. But while interpretable AI/ML can do some of what the cheerleaders for explainable AI/ML desire, adopting a legal requirement of interpretability in automated decision has its own costs and prompts hard tradeoffs: in some cases, the most sophisticated and accurate algorithms cannot be designed as interpretable AI/ML. We argue it is better to face these tradeoffs head on rather than to pretend that explainable AI/ML can provide the kind of explanation we want without tradeoffs. Explainable AI/ML, as currently understood, is therefore an attempt to have our cake and eat it too.

This paper proceeds as follows. Part I provides more background on explainable AI/ML and some of the literature extolling its virtues. It also explains in greater depth the differences between interpretable versus explainable AI/ML. Finally, it provides a synthetic (by which we mean, hypothetical and simplified) example for illustration, which we return to throughout the paper. We will argue that post hoc algorithmic explanations of the form generated by leading explainability algorithms are ineffective along the two dimensions we have described above – they fail to be effectively action-guiding (Part II) and they fail to provide sincere

[21] P.F. Strawson, *Freedom and Resentment*, *in* Freedom and Resentment, and Other Essays 1, 4–6 (1974).
[22] See Zoë A. Johnson King, *Praiseworthy Motivations*, 54 Noûs 408, 409–11 (2020).
[23] *See, e.g.*, Rudin, *supra* note 12.

explanations/motivating reasons for the underlying automated decisions (Part III).[24] In Part IV, we will consider the extent to which interpretability should be a legal requirement, and under what conditions. Finally, in our Conclusion we summarize and also discuss an additional potential problem with explainable AI/ML that we do not fully develop in the main text: namely, that they cannot support reactive attitudes like blame and praise.

I. WHAT IS EXPLAINABLE AI/ML?

Our goal in this Part is for the reader to firmly understand what explainable AI/ML algorithms do and do not do. To understand their limitations, we must first distinguish explainable from interpretable AI/ML.[25] Before we can do this, though, we start with a very general overview of supervised learning. Before we start, a word of comfort: while in the next few sections we use some formal mathematical notation (*X*s, *Y*s, and even some β*s*!), they are not essential for understanding our main arguments – we offer them for readers who want a slightly more technical explanation. Indeed, in Part I.C we use a synthetic example to explain all these points in a more illustrative way.

A. Supervised Learning Models

A typical supervised machine learning or classification model (i.e., a model trained on structured data with labeled features) is effectively a way of solving a function estimation problem using certain optimization techniques. We wish to estimate the response, y (for example, a person's age), as a function of some features, $x_1, \ldots, x_n$ (for example, the person's height and weight). This is a statistical learning task, in part because the way we are going to estimate this relationship is by examining the available data (in our toy example, that would be data on people's ages, heights and weights). Accordingly, we estimate that function by fitting a model to the available data. To fit a model is to solve some optimization problem. For instance, in a typical linear regression model, we have $y = \beta \mathrm{x}^{\mathrm{T}}$, where $\beta \mathrm{x}^{\mathrm{T}}$ is the inner product of a vector of the linear model's parameter coefficients, $(\beta_0, \beta_1, \ldots, \beta_n)$, and the transpose of the vector of input variables, $(x_1, x_2, \ldots, x_n)$. We would then search for the line of best fit, where best is defined in terms of minimizing the sum of squared distances between each point's estimated value and its true value. That is, we choose β so as to minimize $(y - \beta \mathrm{x})^{\mathrm{T}}(y - \beta \mathrm{x})$. For each point, this is the squared difference between y and $\beta \mathrm{x}^{\mathrm{T}}$ (omitting the

[24] Mathilde Cohen, *Sincerity and Reason-Giving: When May Legal Decision Makers Lie*, 59 DEPAUL L. REV. 1091, 1095–96 (2010); Micah Schwartzman, *Judicial Sincerity*, 94 VA. L. REV. 987, 1013–15 (2008); W. Bradley Wendel, *Truthfulness and the Rule of Law*, 35 NOTRE DAME J.L. ETHICS & PUB. POL'Y 795, 816–17 (2021).

[25] *See* Babic et al., *supra* note 13 ("It is important to first distinguish explainable from interpretable AI/ML. These are two very different types of algorithms with different ways of dealing with the problem of opacity . . . .").

subscript i). In more general classification tasks, the basic ingredients are the same: our goal is to identify a function f which will best classify items, on the basis of past observations, where best is defined in terms of minimizing a certain loss function. For each point i, the loss is given by $l(y_i, f(x_i))$. In linear regression, $l(y, f(x)) = (y - f(x))^2$.

It is often the case that the output we are interested in is a probability, and sometimes the ultimate decision is a function of that probability. For example, suppose that we wish to classify a group of people according to their political orientation, and suppose (to make things simple) everyone will be labeled either liberal or conservative. For each person, the algorithm could produce a probability that the person is liberal or conservative. And we could then program the algorithm to make a thresholded decision – namely, we will say the person is liberal/conservative if and only if the predicted probability that they are liberal/conservative exceeds 50%. If there are more than two categories, then we would assign each person to the category that they are most probably predicted to belong to. (With many categories, that could be significantly less than 50%.)

B. Interpretability vs. Explainability

Now suppose we have a classification model given by $y = f(x_1, \ldots, x_n)$ where f is the estimate (model) of the true but unknown underlying function, let's call it g, relating the features $(x_1, \ldots, x_n)$ to the prediction (y). We will say that an AI/ML model f is interpretable (sometimes called intelligible) if an ordinary person can understand how the individual x_i's contribute to the prediction.[26] Let's call this a "white-box" model. The paradigm examples of interpretable or white-box models are linear models or decision trees.[27] In linear regression, it is easy to understand that the predicted y is given by $\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$. This is simple, additive, and generally intuitive. Consider an extremely naive example: we might say that a person's height (in cm) is given by 15 (the y-intercept) + 1.1 × their weight (in lbs). Hence, we would predict that someone who weighs 145 lbs is 15 + 1.1 × 145 = 174.5 cm tall. This works out okay for average values, but is not a particularly good model for low weight individuals. In any case, the idea is extremely simple: take a person's weight, multiply it by some coefficient, and add a fixed "benchmark" value, so to speak. The mechanics behind the prediction are very easy to comprehend.

While this is a helpful way of thinking about interpretability, it is not a perfectly general or objective definition, as it depends on a user's subjective level of expertise or understanding. For example, logistic regression is a type of linear model where the output is a

---

[26] Yin Lou, Rich Caruana & Johannes Gehrke, *Intelligible Models for Classification and Regression*, PROC. 18TH ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING (Aug. 2012), at 1, https://doi.org/10.1145/2339530.2339556 ("By interpretability we mean that users can understand the contribution of individual features in the model.").

[27] Zachary C. Lipton, *The Mythos of Model Interpretability*, 16 ACM QUEUE, May–June 2018, article at 14 ("[E]ach node in a decision tree might correspond to a plain text description. . . . Similarly, the parameters of a linear model could be described as representing strengths of association between each feature and the label [output].").

probability. It uses a link function between the response and the predictors, with the effect that the log odds are linear in the feature variables. Some authors refer to such AI/ML models as interpretable.[28] While not implausible, this assumes quite a lot of statistical understanding on behalf of a user – for example, that the model's linearity is on a log scale, and that it relates the predictors to the probability in odds form.

What this leads us to is the idea that a perfectly general definition of interpretability in the form of a set of necessary and sufficient conditions is probably not possible to provide.[29] Rather, it is better to think of interpretability as existing on a spectrum, whereby some models are obviously opaque black boxes, such as a convolutional neural network[30] with millions of trainable parameters, while others are canonically transparent, such as a linear model with a few predictors or a decision tree with two or three "if then" statements.[31] Often, deciding whether or not an AI/ML model is interpretable is, in the words of Justice Potter Stewart, an instance of "I know it when I see it."[32] An ordinary regression model with one or two variables, such as our example above of predicting a person's weight on the basis of their height, is clearly interpretable. A deep neural network with millions of parameters is clearly not interpretable. But explicitly articulating the boundary of interpretability is no easy feat.

Explainability is very different from interpretability. It does not lie *anywhere* on the interpretability spectrum. Explainable AI/ML attempts to do accomplish the following entirely different task: Given a black-box model f (say, a convolutional neural network), an explainable AI/ML algorithm constructs an interpretable function, h (perhaps a linear model) which approximates f as closely as possible on the available data. In other words, h is a second, separate function, whose goal is to predict as closely as possible to f. To return to our prior notation, we called the true but unknown function relating the inputs to the output g. Explainability is not concerned with approximating g. Rather, the goal of h is to replicate f faithfully, the black-box model.

The idea is that we can use the black box to make the original prediction, and then use the "white-box" approximation of that black box to provide an explanation to a user who requests it. As Cynthia Rudin puts it, "an explanation is a separate model that is supposed to replicate most of the behavior of a black box."[33] Hence, the explainability algorithm explains the black-box *model* (i.e., the

---

[28] *E.g.*, Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor & Mohiuddin Ahmed, *Explainable Artificial Intelligence Approaches: A Survey*, ARXIV (Jan. 23, 2021), http://arxiv.org/abs/2101.09429.

[29] Babic et al., *supra* note 13.

[30] Convolutional neural networks are a type of neural network frequently used in image recognition and similar tasks where the data "has a known grid-like topology." IAN GOODFELLOW ET AL., DEEP LEARNING 321 (2016).

[31] *See* Babic et al., *supra* note 13.

[32] Jacobellis v. Ohio. 378 U.S. 184, 197 (1964) (Stewart, J., concurring).

[33] *See* Rudin, *supra* note 12, preprint at 2.

estimated *f*), by finding a function similar to it (i.e., *h*) and not the underlying *relationship* being modeled (i.e., *g*).

Unlike interpretable AI/ML, explainable AI/ML does *not* attempt to replace the black box with a transparent one. Rather, they aim to approximate the behavior of a black box as closely as possible with a second box which is itself transparent. It follows from this that explainable AI/ML cannot perfectly track a black box's behavior across the full feature space – if they could, then the explainable AI/ML would by definition be equivalent to the black-box model – $f = h$.[34] Hence, this form of explainable AI/ML provides a post hoc rationalization of a black-box prediction. This is the key to our argument and to our critique of the normative value of explainable AI/ML. We will return to this flaw throughout the paper.

Thus far, we have made our claims as to explainable AI/ML *generally*. Now let us zoom in on one particular leading explainability algorithm, LIME (which stands for Local Interpretable Model-Agnostic Explanations), developed by Marco Ribeiro et al.[35] While we focus on LIME to illustrate our arguments, we emphasize that our criticisms are not limited to this algorithm. Our point is more generally about the limited value, from a law and policy perspective, of the *kinds of* explanations that techniques like LIME can produce. Once again, a word of comfort: this is a more technical discussion – for readers not comfortable with mathematical notation they may want to skip it and move straight on to I.C that develops similar points with a more intuitive example.

1. *The LIME Algorithm*

Ribeiro et al. begin with the notion of an interpretable data representation: $x \in R^{d} \longrightarrow x' \in \{0, 1\}^{d'}$.[36] The idea is that the original features (for example, word embeddings) may not be understandable by humans, while interpretable data representations are (for example, whether a certain natural language word is absent or present).[37] Hence, x corresponds to the original vector of features associated with the instance we wish to provide an explanation for, and x' corresponds to a binary vector of its interpretable representation.

Next, we "define an explanation as a model $g \in G$, where G is a class of potentially interpretable models, such as linear models , ... The domain of g is $\{0, 1\}^{d'}$"[38] We also introduce a measure of complexity, $\Omega(g)$, which could for example correspond to the number of meaningful weights in a linear model.[39] We now have our black box model, *f*. Then, we define a notion called locality: $\pi_x(z)$ is "a proximity measure between an instance z to x."[40] This is because we want explanations to

[34] Rudin, *supra* note 12, preprint at 3.
[35] *See generally* Ribeiro et al., *supra* note 13.
[36] *Id.* at 1137.
[37] *Id.*
[38] *Id.*
[39] *Id.*
[40] *Id.*

be *locally* faithful even if they cannot perfectly approximate *f* across the whole space (a concept we might call global faithfulness).

Next, we introduce a penalty for infidelity: $L(f, g, \pi_x)$ is "a measure of how unfaithful g is in approximating f in the locality defined by π_x."[41] The informal idea is then to minimize L and keep $\Omega(g)$ low enough. LIME is then the solution to the following optimization problem:[42]

$$\xi(x) = \ arg\,min_{g \in G} L(f, g, \pi_x) \ + \Omega(g).$$

Now if we assume that G is limited to linear models, and also that L is a proximity weighted square error loss, we obtain:[43]

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi x(z)[f(z) - g(z')]^2$$

A few things to notice from these expressions: An explanation is generated for a particular instance; hence, for different instances it need not be the same explanation. How to select a loss is not something we can answer mathematically. Least squares is a convenient and well understood method, but there are many other options – for example, L1 distance or cross-entropy. We also need to select Ω (the measure of complexity), and π (the measure of fidelity) and this too is not something we can uniquely determine mathematically. Finally, we need to identify G, and this is perhaps the most difficult part of the problem – what is the class of all interpretable models? Ribeiro et al. make a simplifying assumption and treat G as linear models,[44] but this is both over and under inclusive – some interpretable models need not be linear (for example, the classic single layer perceptron), and some linear models can get very complex (for example, a Cox survival model is linear in the log of the parameters, and it can have millions of trainable features). By assuming that G is limited to linear models, Ribeiro et al. effectively show how we can approximate any model with a linear model. They do not shed light on what it takes for a model to be interpretable – they simply take for granted that linear models are interpretable.

C. An Illustration

The prior two sections may have felt abstract or heavy on the mathematical notation. In this section we try to show the same points in a more intuitive way by developing a synthetic example, "The MIT School of Law" – entirely fictitious—whereby a law school is trying to design algorithms to help with its admission criteria. This example is complex enough to illustrate the main problems with current attempts at explainability, but by design an overly simplistic hypothetical approach to the problem of admissions, a sort of "toy" problem.

1. *The MIT School of Law*

[41] *Id.*
[42] *Id.*
[43] *Id.* at 1338.
[44] *Id.* at 1337.

Suppose that having opened its brand new tech savvy law school – we will call it "MIT Law" for short – MIT Law is interested in automating its admissions process. To do this, the school will use admissions data obtained from peer schools for the last 10 years in order to estimate a certain model, which it will then employ to make admissions decisions for its very first class of MIT Law students.

What makes the MIT Law example particularly stylized is that we binarize everything – in our toy problem, there are "admissible students", and "inadmissible students", and all students fall into one category or the other. Information about students (such as their LSAT and GPA) are then used as latent indicators of admissibility. The other reason this example is a little bit fanciful is because it is not clear what "admissibility" means; it is not clear that any such generalized aptitude for law practice exists, and even if it does, it is not clear that it is measurable.[45] At the same time, there does exist something that law schools are trying (very likely imperfectly) to measure, and all that is required for our toy example is the idea that the new MIT Law is trying to do what other law schools are doing by way of an algorithm.

While one *could* try to implement a real-life model like this, and define admissibility in terms of, for example, predicted law school GPA, or in terms of the probability of passing the bar exam, in reality it would probably be a bad idea to assume that all students can be divided as such. Usually there are many competing considerations: we would like to have an intellectually and demographically rounded class, while understanding that different students bring different skills to their cohort.[46] There are also synergistic group dynamics so that the success of a class, however defined, can depend on the composition of the group itself.

Despite its limitations, we will stick with our stylized example because it allows us to vividly illustrate our argument and to describe some of the points in a more tangible way. We do not mean to suggest that there is something particularly important about automating higher education or admission decisions in particular. That said, in the latter parts of this paper we return to the context question – does the kind of explanation that an approach offers matter more for *some* contexts than others – i.e., cancer diagnosis versus sentencing decisions for a criminal offender?

Now, the first question is: what *exactly* is the school interested in predicting? We called it 'admissibility' but what contributes to admissibility? For the sake of our hypothetical, we will assume that MIT Law is willing to be a little bit simplistic and decide who to admit on the basis of their predicted law school performance alone. And we will assume that being extremely quantitative in its approach, MIT

[45]Though, to be fair to our toy model, one could say the same (and many scholars have) about latent measures of aptitude such as IQ, which are also not directly measurable.

[46] To be sure, as current litigation on college admission criteria before the U.S. Supreme Court illustrates, what colleges are and should be measuring is a highly contentious question – but one that we emphasize is not particularly relevant to our paper. *See* Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll., 980 F.3d 157 (1st Cir. 2020), *cert. granted*, 142 S. Ct. 895 (2022). We merely are offering this as an easy to grasp example, nothing more.

Law has decided to look only at applicants' undergraduate GPAs, denoted as x_1, and LSAT scores, denoted as x_2. While simplistic, this is not all that far off from what US News actually has done as a major part of its own rankings of schools,[47] nor is it too far off from how schools evaluate some parts of an application file. We will assume that MIT Law has obtained data on students' past law school performance (perhaps from a peer school), so that for each student in the training set (i.e., the labeled set of structured data to which the model is then fitted before being put to use) they have the student's undergraduate GPAs (x_1) and LSAT scores (x_2), as well as their law school GPAs (y). They now wish to estimate a function $y = f(x_1, x_2)$. For example, if we use a linear model, as described above, we would assume f can be reduced to a function of the form $f(x; \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and we would use the data to identify point estimates of β_0, β_1 and β_2, using the method of least squares (also described above).

We almost have our full model, but a prediction is not a decision. We also need a function which takes the predicted GPA as its input, and, again in our simplified toy example (no wait list!) provides a binary (yes/no) answer regarding admission as its output. The easiest way to accomplish this would be to simply set a threshold on the acceptable predicted law school GPA. For example: admit everyone whose predicted GPA is 3.7 or above. Such a threshold essentially divides our data into two classes – those who are estimated to be admissible and those who are estimated to be inadmissible, where admissibility has been quantified in terms of predicted law school GPA.

2. *Interpretability in Practice*

The model we have just described is very easy to interpret. If a student is rejected and inquires as to the nature of her decision, the school can provide a simple, understandable, and transparent explanation. The predicted output is simply a sum of the inputs weighted by their parameter estimates. The model uses a weighted combination of an applicant's GPA and LSAT score, and nothing else. We can make the model's parameters available, which would then enable the applicant or anyone else to determine the feasible combinations of LSAT and GPA that he or she likely needs to be admitted the following year.

This information is not just "nice to have," but crucial in terms of the kinds of explanations that applicants want and, we would argue, a more general desideratum of explanations in algorithmic decision making. A transparent model is suitably *action guiding* – it is actually useful in shaping the student's subsequent behavior and

[47] Robert Morse et al., *Methodology: 2023 Best Law Schools Rankings*, U.S. NEWS (Mar 28, 2022), https://www.usnews.com/education/best-graduate-schools/articles/law-schools-methodology; *see also* Robert Morse & Stephanie Salmon, *Plans for Publication of the 2023-2024 Best Law Schools*, U.S. NEWS (Jan. 2, 2023), https://www.usnews.com/education/blogs/college-rankings-blog/articles/2023-01-02/plans-for-publication-of-the-2023-2024-best-law-schools.

helping them plan for the future. If a student has his or her heart set on MIT Law and wants to re-apply, should the student study for the LSAT more? Should he or she instead take some additional undergraduate courses in order to improve his or her GPA? These questions can be meaningfully answered if the student is given the transparent model that was used to evaluate their application.[48]

Now suppose that the data MIT Law has obtained looks as follows.

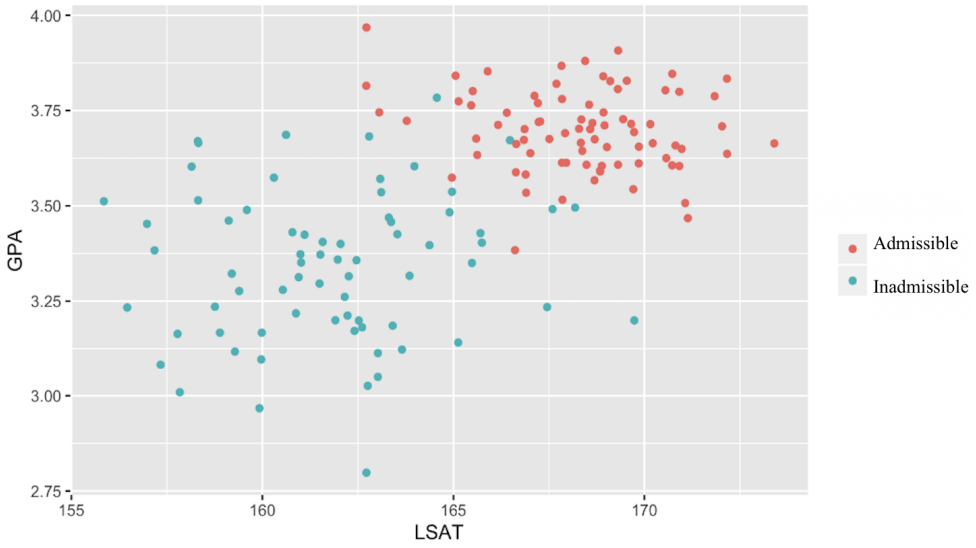

Figure 1. Hypothetical law school data.

Figure 1 depicts a synthetic data set that we have created for 150 students. As described, we have (again, made up!) information about students' LSAT scores and undergraduate GPAs. We assume there are two types of students: admissible and inadmissible. This is just a generalization of our 3.7 threshold from above.

We generated this hypothetical dataset by assuming that admissible students have on average 168 LSAT and 3.7 college GPA whereas inadmissible students have on average 162 LSAT and 3.3 college GPA, where both groups are normally distributed. What this means is that admissible students do better, on average, in the observed features, but there are still admissible students who just so happen to obtain a low undergraduate GPA or low LSAT score (for example: a student may get sick on the day of the test, or do poorly in an undergraduate semester due to unexpected health problems) and inadmissible students who happen to obtain a high undergraduate GPA and LSAT score (for example, due to luck or due to taking

[48] While it is true that the pool for the following year will always look different, this transparent model can precisely at least answer the question "what would I have had to do differently to have been admitted this year," and that is very helpful in guiding their actions for the next application cycle.

particularly easy courses whose grade is not reflective of their aptitude).

MIT Law's task now is to identify a model that best distinguishes the two groups. We can make estimates using different models. The panels below show two attempts: on the left, we have a linear model (the kind that we described above) and on the right we have a single layer feed-forward neural network. Both have been fit to the hypothetical law school data.

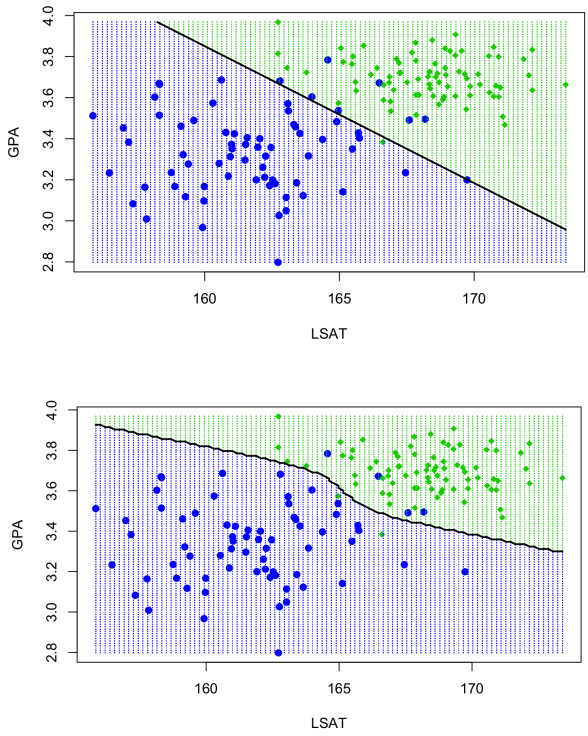

Figure 2. A linear model (top) and neural network (bottom) fit to law school data.

The black line in each panel is the fitted model – this is the classification boundary. When a new student applies to law school, we would take their LSAT and college GPA and plot them on this grid. If they are below the line they would be rejected. If they are on or above the line, they would be accepted. Notice that the simple model (left) does just a little bit worse than the neural network (right). By worse we mean that the model "incorrectly" admits 8 inadmissible students in the test data, and "incorrectly" rejects 1 admissible student (9/150 mistakes). The neural network contours better around the groups, so that it "incorrectly" accepts only 4 students and "incorrectly" rejects 2 (6/150 mistakes).

The linear model on the left can be described as before, using an expression of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. The prediction is a simple linear function of one's LSAT and undergraduate GPA. But the neural network model cannot be described this way. Instead, the (very

simple) neural network model we have used takes a student's LSAT and GPA, connects them to more than a dozen nodes in a hidden layer, and then connects those nodes to the outcome. The result of this process is that it would be hard to describe "how much" of a role one's LSAT plays, and how much of a role one's GPA plays. It would also be hard to know how the relative importance of those variables changes as we update the model. It would not be possible for a rejected applicant to grasp the feasible combinations of LSAT and GPA that would lead them to be admitted the following year. If she retakes the LSAT, how much better does she need to do? It is hard to say. The only way for her to really answer this question under various hypothetical scenarios is to run this model on her own computer, if she was given access, and feed it various possible combinations of LSAT and GPA and evaluate the outcome.[49]

In the hypothetical case we are considering, one thing we can do with both the linear model and the neural network model is to show to all applicants the classification boundaries in Figure 2. That can be helpful to guide action, but note that those boundaries change as the model is updated with new data (more applicants, more graduates), so what we would really like to understand is how the variables are combining to formulate each prediction. Moreover, visualizing the classification boundary is only possible in toy cases where the number of variables is three or less (in our case we have two). In real life, a good model might have tens or hundreds or thousands of variables depending on the question. In general, for an *n*-dimensional classification problem, the classification boundary is an *n-1* dimensional hyperplane. In our example, it is a line segment. With three variables, it would be a two-dimensional slice through a three-dimensional plot. Beyond that, we can no longer rely on visualization in the same way.

Thus far we have contrasted a simple linear model, which is interpretable, with a neural net (NN) model, which is not. We have not yet said anything about *explainable* AI/ML in this example. As we described above, leading explainable AI/ML, such as LIME[50] and SHAP,[51] generate a supposedly transparent "white-box" model that tries to explain a non-interpretable "black-box" AI/ML model. That sounds good, but what does that actually mean in practice? Let us explain by imagining how MIT Law might do exactly that.

[49] Coincidentally, this is indeed how some authors have tried to articulate what explainability means – namely, to allow users to interact and run the model even if they do not understand its internal workings. *See* Lipton, *supra* note 27, article at 15. We are not in principle opposed to this practice, it can provide limited benefits. But our argument is more narrowly directed at leading explainability algorithms.

[50] Ribeiro et al., *supra* note 13.

[51] Lundberg & Lee, *supra* note 13.

3. *Explainability in Practice*

Suppose that after constructing the NN model in the right panel of Figure 2, we then engineer a linear model that locally approximates the NN model as closely as possible, as in Figure 3, below.

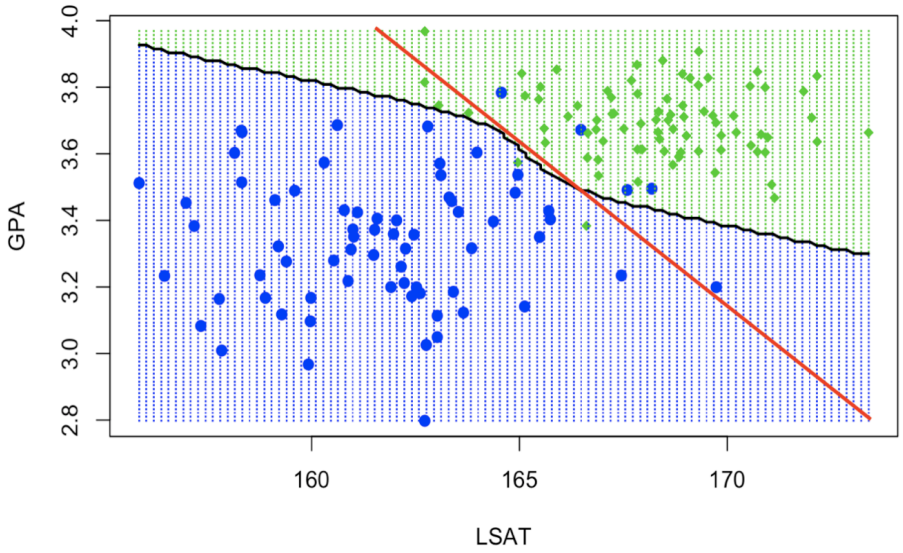

Figure 3. Linear approximation (red line segment) of NN model (black curve) fit to law school data.

The model depicted by the red line is of the simple linear form. We can therefore use that model (the red line segment) in order to explain the uninterpretable one (black curve) for a particular student whose LSAT and GPA land them in the neighborhood where the red line segment is approximately tangent to the black curve. If a student asks why she was rejected from MIT Law and the school wanted to answer truthfully (if somewhat technical in its form, although it is MIT after all!) it could tell her: a linear approximation of our black-box model suggests that this is roughly the formula it applied to your application file. It is not the *actual* formula used, but it is our best simplifying guess.

The student can then (supposedly, according to explainable AI/ML proponents) use that formula in order to guide her subsequent behavior – for example, to estimate how much better she needs to do on the LSAT the next time around to get admitted to MIT Law. This is what explainable AI/ML algorithms *attempt* to do – to give an explanation for a non-interpretable AI/ML model. In this case we have offered an illustration of the influential LIME[52] explainability algorithm. It is this kind of explainability algorithm that is the subject

[52] Ribeiro et al., *supra* note 13, at 1135; *see also* Lundberg & Lee, *supra* note 13, at 2 (similar SHAP algorithm).

of our arguments – algorithms, like LIME, which produce post hoc rationales of black box model predictions.[53] In the next section, we explain what it is about the post hoc nature of those rationales that make them unable to do what people want explanations to do.

II. WHY EXPLAINABLE AI/ML CANNOT ACHIEVE ITS GOAL OF ACTION GUIDANCE

As we have discussed in Part I, an interpretable AI/ML model is a model that can be understood immediately. Explainable AI/ML, by contrast, is one for which we can construct a secondary approximating model which can itself be understood. While there are many different approaches to explanation, the linear approximation is a paradigmatic example, and they all share a family resemblance in the sense that we try to glean insights about what is happening inside the black box without opening it up. Meanwhile, with an interpretable AI/ML model, we simply avoid using the black box in the first place.

In this Part, we demonstrate why explainable AI/ML models cannot do the work that their proponents would like them to do. Specifically: They fail to effectively guide action, they fail to provide sincere explanations/motivating reasons for the underlying automated decisions, and they cannot underwrite normative attitudes like blame and praise. We develop each point in turn.

A. Effective Action Guidance

If a student has been rejected from law school in our example, and she seeks an explanation, this could be either because she would like to request a review of the decision or, more likely, because she would like to know what she needs to change in order to have a better chance at being admitted next year. Similarly, if a loan applicant is denied a loan, she wants to know what she needs to change in order to have a better shot at being given a loan by the next institution or the next time around.[54] If a defendant is denied parole, she wants to know how to change her behavior in prison in order to have a better shot at the next hearing.[55] If an applicant is not hired, she wants to know which skill sets to cultivate in order to have a better chance next recruitment season, and so forth.[56] The point is that explanations are valuable in no small part because they can be action guiding. A causal explanation, for instance, is valuable because it provides understanding about which input has to change in order to change the output.

Consider a non-algorithmic analogy to show you how this is related to explainability. Suppose you are waiting for a friend, Dave, to

---

[53] We understand that "explainable AI" is often used much more broadly, to refer to any kind of system that sheds some light on an algorithmic prediction. But in this project we limit our attention to LIME and its counterparts, which are indeed the leading explainability algorithms.
[54] *See* Magnuson, *supra* note 1, at 349–50.
[55] *See* Coglianese & Ben Dor, *supra* note 1, at 802–04.
[56] *See* Ajunwa, *supra* note 1, at 623.

meet you at a movie theater. After waiting for 30 minutes you call Dave's partner, Sidney, to ask if they know where Dave is because you want some guidance between whether to (a) buy tickets and snacks for you and Dave because Dave will be along in a minute, (b) get tickets for a later show because Dave is going to miss this one, or (c) see the show you originally intended alone since Dave is not coming at all today. Sidney says: "Hmm. Dave could be delayed due to a work emergency." You thank Sidney for the information, but in truth this explanation is not very helpful to you. *If* there were a reason to believe Dave *had* a work emergency, that would be useful to you in deciding among your options; but simply suggesting that there *could* be a work emergency because a work emergency is consistent with "the data" (i.e., Dave being late) is of little use. If Dave had a work emergency you might pursue option (b), if someone in Dave's family got sick and needs his help you may pursue option (c), and if Dave is just stuck in traffic you might pursue option (a). The explanation Sidney has given you does not tell you which of these it is, all could be the explanation, and for that reason her advice is not action guiding.

Now let us return to algorithmic models and their explanation. For similar reasons, explainable AI/ML is not nearly as useful in its action guidance, nor is it as valuable in its transparency, as an interpretable model can be. For example, if the student who was rejected by MIT Law attempts to use the explainable AI/ML model in our hypothetical to guide her subsequent behavior, the relevance of the approximation depends on where the student falls. Applicants near the mean of both groups (around 165 LSAT and 3.6 GPA) are classified very similarly by both the interpretable and explainable AI/ML models illustrated above. Indeed, in the neighborhood of the mean values, the only way to be classified differently by the original model and its explaining model (i.e., to be admitted by the original and rejected by the explanatory, or vice versa) would be to fall exactly in the vanishingly small space between the black curve and red line segment in Figure 3.

Meanwhile, applicants at the two extremes can be classified differently (i.e., receive a different result) by the different models very easily. For example, for students with an LSAT score of 175, everyone with a GPA between approximately 3 and 3.4 will be treated differently by the actual model and its linear approximation. To understand why, we direct the reader to Figure 3, above. Notice that around LSAT scores of 165, the linear approximation (red line segment) is approximately secant to the original neural network (black curve) – i.e., in that small region it is a very good approximation, and there is almost no space between them where a point could fall. But, around LSAT scores of 175, the space between the red line segment and the black curve is very large. Every applicant who falls into the space between them would be classified differently by the original model and its approximation. So for these students the linear approximation is neither revealing of what is happening inside the black box nor very useful in guiding their future behavior. Indeed, if they attempt to use it for action guidance, it will severely mislead

them. A student with a 175 LSAT score would assume she only needs approximately 3.0 GPA for admission when in reality she needs at least 3.4.

More generally, the point is that when our rejected law school applicant learns that she will be evaluated using a point system that is based, additively, on her LSAT score and her GPA – as she would, if the admissions system were based on a simple interpretable AI/ML model – she can plan for the future. She can, for example, consider how much time she has to study for the LSAT, take a prep course, etc., if it seems feasible that this will lead to her admission. If her LSAT score was already near perfect, she could discern that a marginal improvement will not change the result. If her GPA was already high, she may want instead to spend more time improving her LSAT score. And so forth. She knows which knobs she needs to turn, and by how much. This is exactly what she wants, as does our loan or parole applicant in those examples.

But the explanations generated by explainability algorithms fail to provide this kind of action guidance, precisely because they do not reveal the actual mechanism by which the original decision is made. They tell the student, in effect: it could be that you were rejected because your LSAT score was too low, as this would be consistent with the data. But despite this explanation, it could *also* be that the student was rejected because her GPA was too low, or because their combination was too low, etc. Our rejected student would not learn anything about the admissions procedure from a post hoc explanation that is consistent with the data. The post hoc explanation is simply an explanation that is permitted given the data, because it is not inconsistent with it.

Proponents of explainable AI/ML may point out that such explanations are *local*, meaning that for every given applicant we generate a unique explanation (i.e., a unique red line segment consistent with their feature values). But this reinforces our point about the post hoc nature of the explanation and its inability to guide future action. If a unique explanation is generated for every applicant, then we can no longer even pretend to be shedding light on the actual classification boundary that was applied. We elaborate on this point below.

III. WHY EXPLAINABLE AI/ML IS INSINCERE (AND WHY IT MATTERS)

Explainability models like LIME ordinarily generate a unique explanation for every instance. That is: we take an instance (a law school applicant, in our example), and given the decision that they did receive (admit/reject), we generate a plausible linear model that could have produced that decision. In other words, algorithms like LIME generate a different explanatory model for every instance. Given this, one might take issue with our discussion above – where we treat the approximating model (the red line segment in Figure 4) as fixed, and consider how different students would fare had that model been

applied to their case. In reality, as we noted above, every student would be shown a slightly different red line segment that is consistent with their case, because the loss function that we are minimizing includes a term that pertains to proximity from a particular instance (as we explained in I.B.1).

However, far from being a saving grace for AI/ML explainability models like LIME, this poses an additional problem for them. Instead of producing a stable and robust explanation, it becomes clear that the explanation they produce is a post hoc rationalization as soon as one realizes that it can differ from instance to instance (or from applicant to applicant, in our example). Otherwise put, explainable AI/ML is explicitly and inherently *insincere* about the grounds for a decision. In fact, even for a single applicant we can generate more than one explanation.

A. Illustrating Explainable AI/ML's Insincerity

An easy way to illustrate explainable AI/ML's insincerity is to consider further how the model might generate different explanations for different people. Before we return to MIT Law, let us illustrate why this is a problem with an even simpler example. Imagine you are a man going on a date with someone. The dater ends the date by telling you: "you are amazing, you are exactly the kind of person I want to date except I won't date men under six feet. I am so sorry. I am sure you will find someone great." Your feelings are hurt, but at least you believe it really was your height, which is not something you can control. A month later, you discover the dater is seriously dating someone who is 5'8". You might have many thoughts about the dater and the person's dating behavior, but one immediate feeling you might have is that the dater has been insincere. The explanation given was not *the* explanation (even though in your case it was consistent with "the data" at that point in time) – because if it had been *the* explanation, the dater would not be with the other chap either.

Now let us return to MIT Law and see how the same is true there. The explainable AI/ML model generates inherently insincere explanations as evidenced by the fact that it would generate different explanations for different applicants, since not all applicants are classified the same way by the two models (the black box and its approximating white box). Accordingly, by hypothesis, there exists some pair of applicants that would receive different explanations. Indeed, this will be true for many pairs of applicants. Consider, for example, the student in our model who is rejected with a 170 LSAT score (bottom right oval in Figure 4). The approximating model we have been using to illustrate our argument so far would admit her, so if we want to explain why she was rejected we need a second approximation, as in Figure 4, below.

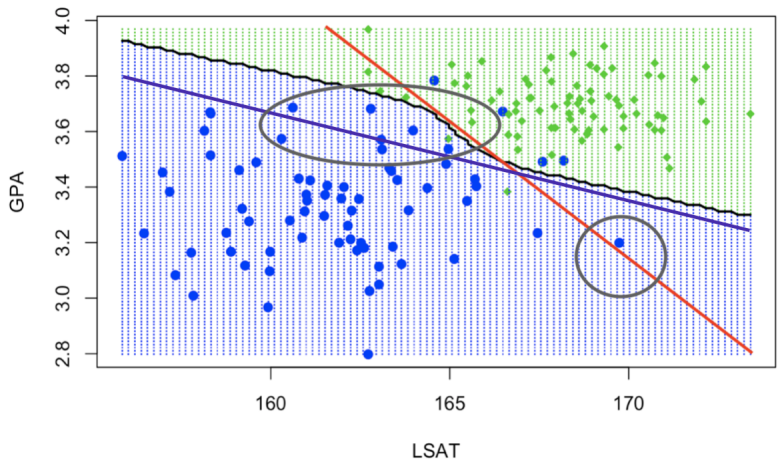

Figure 4. Competing linear approximations of neural net model fit to law school data.

The purple line segment offers a second, competing linear approximation of the neural net model (the black curve). For the student corresponding to the point inside the bottom right rounded oval, this model would count as a more effective explanation since it is now consistent with the NN model (unlike the red line segment, which is inconsistent with the original model for this student, because it would admit the student whereas the NN model would reject the student).

This reinforces what we have been calling the post hoc nature of algorithmically generated explanations. We look at where the student falls in the original model, and we identify an "explanation" that reinforces the original decision. For the same reason, multiple competing explanations can be mutually inconsistent across a group of instances (applicants). If we use the purple line segment as our explanation, then the students inside the oval on the top left should be admitted. But if we use the red line segment as our explanation, then the student on the bottom right should be admitted. But under neither line segment are they admitted or rejected together – that is, neither explanation can predict what the model tells us to do for *both* students simultaneously. And if these students can communicate with each other, then they will know we are being insincere to at least one of them – that is, the explanation we gave one of them does not apply "the rule" because it does not explain the result that occurred to the other. Thus, far from promoting trust and transparency, our hypothetical MIT Law will be almost sure to *undermine* its applicants' trust by using an explainability algorithm in its admissions process.

In the examples we have been using there has been a feature that has exposed the insincerity. In our dating hypothetical, our rejected man only learned about the insincerity of the dater's explanation because of his observing the dater's subsequent relationship. Likewise, our applicant to MIT Law only knows about the insincerity of the school's algorithm by comparing notes with the

other applicants and their own queries and responses. In many instances, those who are adversely affected by an algorithm will not have as ready an opportunity to "share and compare" and thus detect the insincerity.

But that is *not* a mark *in favo*r of the explainable AI/ML algorithm. Indeed, if one were to adopt an explainable AI/ML algorithm precisely because it makes the detection of insincerity more difficult, that would be a mark *against* it for a system designer whose argument for adopting the algorithm is to be transparent and promote trust. It would be particularly noxious to tout the benefit of one's algorithm as giving transparent explanations when those explanations are insincere and contradictory, but those features are hard to detect. Insincere explanations are, in many instances, not the kinds of explanations worth wanting.

So far, we have focused on the LIME algorithm for illustration because of its ubiquity and simplicity, but the problem is not limited to the LIME algorithm. Consider another leading algorithm developed by Lundberg and Lee.[57] It is called SHAP (Shapley Additive Explanation). The idea is based on a solution concept from cooperative game theory known as a Shapley value.[58] Like LIME, this algorithm identifies a model which estimates the feature importance of a black-box model. In the context of explanation, the Shapley value provides the average contribution of a feature. In this sense, both LIME and SHAP are what Lundberg and Lee call "additive feature attribution methods."[59] In our simple law school admissions example, they are linear approximations without interaction (they enable us to identify the appropriate red line so to speak, as in Figure 3). They construct a model that allows us to make statements such as those we would make in the MIT Law example: by simply adding up the feature contributions we approximate the model's prediction. Like LIME, SHAP is a post hoc exercise, and like LIME it can only be guaranteed to be locally faithful. It cannot be guaranteed to be an accurate approximation everywhere because, again, that would mean it is equivalent to the original model and in that case we would no longer need the original model.

B. Why Insincere Explanations Are a Problem

We have argued that instead of providing the actual reasons for a decision, Explainable AI offers post hoc rationalizations. Some of what is distasteful about such explanations came out in the dating example, and the more general MIT Law hypothetical, but it is worth spending some time to more formally ask whether and why post hoc rationalizations are actually bad. Otherwise put, what makes the insincerity of post hoc rationalizations a problem?

---

[57] Lundberg & Lee, *supra* note 13.

[58] L.S. Shapley, *Notes on the n-Person Game – II: The Value of an n-Person Game*, RAND CORP. WORKING PAPER DOC. NO. RM-670-PR (Aug. 21, 1951), https://doi.org/10.7249/RM0670.

[59] Lundberg & Lee, *supra* note 13, at 2.

Consider the kinds of algorithms, such as COMPAS,[60] that we have seen used in the criminal justice system in ways that have been heavily criticized.[61] Once again we will just describe it in an intentionally oversimplified way to illustrate the point. Start with an analogy. Suppose that a certain defendant is denied parole one morning by a presiding judge or probation officer. This defendant wants to know why they were denied parole, and asks the judge's clerk. After the judge has made her decisions for that session, the clerk takes a look at that morning's data, and searches for a pattern. Conveniently, it turns out, everyone granted parole that morning has been a volunteer of the prison book club. The defendant in question was not a volunteer of the prison book club. Hence the clerk tells the defendant: the reason for your denial of parole is that unlike all those granted parole this morning, you did not volunteer in the prison book club.

In what way was the clerk's explanation insincere? Because of its post hoc and contingent nature. The clerk picked out this explanation because the clerk recognized that the defendant was not a volunteer of the prison book club and, coincidentally, everyone granted parole was a book club volunteer. The clerk did not pick out this explanation due to a sincere belief that the judge's conclusion is causally determined by membership in the prison book club. This explanation therefore is neither *unique* nor the *actual* reason for the judge's decision. It is simply one of many patterns that the clerk was able to identify after the fact in order to justify or rationalize the judge's decision.

Now here is the crux of the point: explainable AI/ML behaves much like the clerk in our analogy. And insofar as one finds the clerk's insincerity to be objectionable, then Explainable AI/ML's insincerity, as exemplified by algorithms like LIME and SHAP, is objectionable in a similar way.

While that sounds intuitive, we should be careful to examine in a little bit more detail what is meant by insincerity to understand why it is bad. First, there is a sense in which such insincerity is self-evidently a bad practice, and a sense in which it makes explainable AI self-undermining. We know, by hypothesis, that the reason given is *not* in fact the *actual* reason for this defendant's denial of parole – indeed, if we expand our observations and look at decisions made on other days we would learn that many people who were not volunteers for the prison book club but who had otherwise stellar behavior records *did* receive parole.[62] Hence, the rationale is false.

---

[60] Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV (Nov. 17, 2016), https://arxiv.org/abs/1609.05807 (indicating Jan. 2017 presentation at 8th Innovations in Theoretical Computer Science conference).

[61] *See* Kay Firth-Butterfield, *Artificial Intelligence and the Law: More Questions than Answers?*, 14 SCITECH LAWYER 28 (2017); Sophie Noiret et al., *Bias and Fairness in Computer Vision Applications of the Criminal Justice System*, 2021 IEEE SYMP. SERIES COMPUTATIONAL INTELL.; Gijs van Dijck, *Predicting Recidivism Risk Meets AI Act*, 28 EUR. J. CRIM. POL'Y & RSCH. 407 (2022).

[62] Of course, it would be different if there really was a causal relationship between the book club and receiving parole. In our hypothetical example that is not what we intend. And indeed it would be quite a bad policy for criminal justice if this was *the* reason.

The explanation is thereby a kind of fool's gold, so to speak, or a kind of moral sedative. It is provided merely to placate the defendant by providing a plausible rationale of why the judge might have decided the way she did even though we know that is not in fact *the* reason for the judge's decision. If a public interest litigation group wanted to challenge the judge's decision making, it might wrongfully spend time trying to put pressure on the judge's book club practice, when in fact that is not *the* explanation for the result. If criminal defendants – or perhaps more likely, a group of criminal defense lawyers or members of the public defender's office – started talking amongst each other, then their faith in the explanations given would also crumble when they realize that the explanations are just insincere post hoc rationalizations. More generally, once users of algorithms become aware that they are receiving explanations which are false in this sense, that will undermine the system's credibility, as well as the user's trust in the system (which is antithetical to what explanations attempt to accomplish in the first place).

What would be a better form of explanation in this case? It would be normatively more desirable to say that while we do not know why the judge decided the way she did in this particular case, her decision is defensible on multiple grounds, and among them is that volunteering in the prison book club might have improved the defendant's overall score. This would at least be an honest assessment. This is valuable because we often run the risk of uncovering patterns with post hoc explanations which we know ex ante not to be causally relevant. For example, suppose the clerk instead recognized that everyone granted parole on the relevant morning was wearing a blue t-shirt. In this situation, it would be more harmful to the system's legitimacy to produce this as *the* explanation, than to simply produce nothing. Why? Because we have good reason to believe that wearing a blue t-shirt would not, and *should* not, improve the defendant's score (indeed, that is not something the judge should be looking at, at all). But when algorithms generate explanations, they cannot distinguish between the book club story and the blue t-shirt story. The explaining algorithm does not have a causal picture of what is happening in the original black-box algorithm. The reason we know that the latter is patently false is because we are using our prior information about how the judge might be reasoning, and we know that t-shirt color should be legally irrelevant; hence any such association most probably occurred by chance. It is far less clear ex ante that prison volunteer work will be causally irrelevant, and so if offered this explanation one might mistakenly take it as *the* explanation and thus action guiding.

What this discussion illustrates is that while post hoc explanations are ordinarily arbitrary in the sense of not producing the actual (causally relevant) reason for a decision, some such explanations can be legitimate, or legally justifiable (the book club explanation), while others are not legitimate (the t-shirt color explanation). In the latter camp we could also include explanations based on gender, ethnicity, or race, for example. When the explanation is not justifiable, that is an immediate problem – because for

normative/policy reasons we do not want judges to rely on features such as, say, race or ethnicity. When the explanation is justifiable, that is better, but there can still be a problem – which is the action guiding dimension that we have discussed. Even the book club explanation is a bad explanation, because we do not know how if at all joining the book club may affect this defendant's chances later on. If we knew that as a general matter joining the prison book club boosts one's behavior score, then the book club explanation would be effective. But we do not know this.

We hope we have explained why the failure of explainable AI/ML to effectively guide action and its tendency to generate insincere explanations are problems. But to make matters worse, they are problems that run counter to the very benefits of explainable AI/ML that its proponents champion. For example, Ribeiro et al. write that: "Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model."[63] Similarly, Lundberg and Lee start their paper by observing: "The ability to correctly interpret a prediction model's output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled."[64]

From a behavioral perspective, however, it is hard to see how an insincere explanation of this sort could provide the immediate benefits these authors tout, once its users recognize the insincerity. Would this increase trust in the parole process? Maybe in a one-shot round, so to speak, but certainly not in the long run. Does it increase the transparency of the judge's reasoning? Not at all. In fact it often obscures it. Does it promote democratic accountability and the rule of law? Doubtfully – on the contrary, picking out an insincere rationalization and pretending it to be *the* reason for a decision seems to undermine the standing and legitimacy of the judicial office. But this is exactly what Explainable AI/ML does. We apply a black-box model to generate a prediction. We don't know the actual reasons for that prediction. We then approximate the model using a white-box, and we give the reasons associated with the white-box prediction, pretending them to be the *actual* reasons.

C. An Objection: Sincerity vs. Justification

We have given a fairly intuitive argument for why insincere explanations are a problem. But it is worthwhile to pause to consider whether the argument, though intuitive, might be wrong. Some scholars have recently suggested that what matters more for

[63] Ribeiro et al., *supra* note 13, at 1135.
[64] Lundberg & Lee, *supra* note 13, at 1.

procedural integrity is not so much sincerity, but rather justification.[65] When we ask whether an algorithmic decision can be justified we want reasons, Gillian Hadfield argues, that "follow the rules of our community."[66] But those reasons, she argues, need not be unique, nor do they need to be the actual reasons for a decision.

Consider a lending example, which Hadfield raises:[67] an applicant is denied a loan by a bank and seeks to know why she was denied the loan. The bank has a black-box algorithm which deemed the applicant too risky to lend to. While the bank may not know the actual feature values that led to the denial, Hadfield argues that it can give any legitimate explanation consistent with the data. For example, it can say: your credit score was not high enough, your annual income was not high enough, your work experience is insubstantial, etc. These would all presumably be true in such a case. And any of these is a legitimate justification. By comparison, an illegitimate justification would be: you belong to X minority race and we decided not to lend to applicants of this race. So she argues that legitimate reasons – those that will make the decision justifiable – need not be the *actual* reasons. And on Hadfield's view, while procedural justice requires legitimacy (and in turn justifiability) it does *not* require sincerity. We simply have to produce s*ome* legitimate reasons – not necessarily the ones that causally brought about the decision maker's judgment.

This position echoes some writing in legal scholarship outside the context of algorithmic decision making. Mathilde Cohen, for example, argues that lack of sincere reason giving is in many legal contexts not an impediment to the legal validity of a decision.[68] To make this argument, she first distinguishes between motivating and normative reasons, following the renowned philosophers Bernard Williams and Thomas Nagel.[69] Motivating reasons are said to explain a person's actions – they are the reason in virtue of which the action was taken.[70] Normative reasons, on the other hand, are said to justify a person's action rather than explaining it.[71] For instance, in the lending example, the fact that a denied applicant's credit score is insufficiently high can be a normative reason without being a motivating reason. By normative reason, Cohen (following Williams and Nagel, among others) means that it is a reason which, in the context of our legal institutions, could constitute a legitimate ground for the decision.[72] But it may not be the actual ground for this

[65] Gillian K. Hadfield, *Explanation and Justification: AI Decision-Making, Law, and the Rights of Citizens*, UNIV. OF TORONTO SCHWARTZ REISMAN INST. FOR TECH. & SOC'Y (May 18, 2021), https://srinstitute.utoronto.ca/news/hadfield-justifiable-ai.
[66] *Id.*
[67] *Id.*
[68] Cohen, *supra* note 24, at 1115–21.
[69] *See id.* at 1097 n.23 (citing BERNARD WILLIAMS, *Internal and External Reasons, in* MORAL LUCK**:** PHILOSOPHICAL PAPERS 1973–1980 101 (1981)); *see also* B.A.O. Williams & T. Nagel, *Moral Luck*, 50 (Supplementary Volume) PROC. ARISTOTELIAN SOCIETY 115 (1976).
[70] Cohen, *supra* note 24, at 1097 n.23.
[71] *Id.*
[72] *See id*. at 1107–08.

particular judge's decision.[73] Meanwhile, the fact that the applicant belongs to a minority race could be a motivating reason without being a normative reason. That is, it could be the actual ground for this particular judge's decision even though the judge *should not* be deciding on that ground.

Having made this distinction, Cohen then argues that there are two ways we can understand sincerity: as a requirement that our stated motivating reasons correspond to the reasons that in fact motivated us (internalist reading) or as a requirement that our stated normative reasons correspond to reasons which are in fact legitimate justificatory reasons (externalist reading).[74] Finally, Cohen then argues that while externalist sincerity can sometimes be called for, internalist sincerity is rarely a jurisprudential requirement.[75] Hence, as Hadfield suggests, when a loan seeker's application is denied, the decision maker need merely to present normative reasons – i.e., reasons that justify the decision. Those reasons need not be the reasons that actually motivated the decision maker.

While this is an interesting and compelling argument, we think there is a bit of a sleight of hand here, and that sleight of hand is further exacerbated if we try to extend Cohen's argument to the algorithmic context (she does not do this) and to argue on its basis that sincerity should not be a requirement for explainable AI/ML. The problem for her argument is because the ordinary notion of sincerity is inextricably bound up with what Cohen would call internalist reasons. The social practice of reason giving in the common law legal tradition is thought of as giving not just a reason, but a *motivating* reason.

Consider our dating example again: after the dater explains they do not date anyone under 6 feet, you spot them with someone who is clearly well below that height. Feeling like you've been lied to and misled, you ask the dater why they were insincere about their rationale for not dating you. The dater says: "I was not insincere. Height is one among many normatively permissible reasons to reject a dating prospect. While it was not *my* reason, it is indeed a justifiable reason given our dating norms and practices. Hence I gave you a sincere explanation, albeit on an externalist reading of sincerity." At this point, we suspect, you would believe that the dater is trying to be a little too clever with you. All you want to know is *why* they did not want to date you. It does not help for the dater to report one among many *possible "permissible"* reasons for why one person does not like another enough to date them, even though this was not *the dater's* reason. That is simply not what you are interested in, and if that was what sincerity required us to give, sincerity would cease to be of any real value to us.

The value of sincerity is connected to action guidance, the topic of Part II. Internalist sincerity, and the associated motivating reasons it requires decision makers to produce, is valuable not just in and of itself, but in order for explanations to play their action guiding role.

[73] *See id*. at 1097.
[74] *Id.* at 1122.
[75] *See id.* at 1137. *But see id.* at 1138 (suggesting a context-sensitive approach).

Consider a different example. Suppose we have two navy aircraft pilots, whose task is often to fly in sequence, one behind the other (such as in reconnaissance missions). Call the front pilot Maverick and the trailing pilot behind him Goose. It is important for Maverick and Goose to correspond some of their behaviors to each other. In particular, it is important for Goose to understand when and why Maverick slows down so that Goose can react accordingly. Part of reacting accordingly is to predict how Maverick will react to various conditions – in some instances Goose will not have time to wait and see what Maverick does, and if he does not not start slowing down early enough the two may crash.

Now, for any given instance in which the Maverick slows down, it is possible that there are multiple justifiable reasons that he could produce for taking that action. For example: Maverick saw something noteworthy on the ground, the cloud cover became too thick, the distance between Maverick and Goose had become too large, and so forth. These are all legitimate reasons to slow down. But Goose, flying behind Maverick, needs to know the *actual* reason Maverick is slowing down. If the Maverick says "I slowed down because the distance between us grew too large," and that is not the motivating reason, then the next time the distance grows too large, Goose will slow down while Maverick may not. Even more worrisome, if in fact Maverick's motivating reason is "I slowed down because the cloud cover became too thick," but Goose mistakenly thinks this is not Maverick's motivating reason, then the next time the cloud cover gets at least as thick Goose will not slow down while Maverick does and they may crash. It doesn't matter that both of these explanations are perfectly *justifiable*. What is needed to guide Goose's behavior is the motivating reason. Otherwise put, the fact that Maverick is inconsistent with their reason giving is a problem for action guidance – *even if* all the reasons they might give are permissible normative reasons.

For algorithmically generated decisions to be appropriately action guiding we think the same is true. It is not enough that the reasons be justifiable, i.e., normatively permissible. What is needed for true action guidance is that the reasons given be the motivating reasons, i.e., sincere reasons.

Indeed, not only are insincere reasons not helpful, but they may be pernicious. To illustrate, consider another toy hypothetical involving sentencing. Imagine that in every sentencing hearing involving a black defendant a particular judge offers a pretextual (internally insincere but normatively sincere) reason for the decision, which covers up the judge's racism underlying the severity of the sentence given. For example, the judge offers as a reason "the higher sentence is warranted because the accused has failed to show any remorse over the crime." Suppose that is a normatively justified reason that it is perfectly appropriate as a motivating reason for the

judge to rely upon.[76] But it turns out that while it is an appropriate justification, it is one the judge never deploys in cases involving the sentencing of white defendants. This would then lead to a divergent set of sentences across racial groups that seems problematic *even if* in every individual case involving a black defendant there was a post-hoc rationalization that could be justifiably given. The discrimination may not be evident in the one-shot case, but over time the pattern reveals itself in the judge's inconsistency. That is somewhat like what happens with our MIT Law example when one explanation is given to one MIT applicant and another to a different one.

In short: if we try to extend Cohen's argument that sincerity should not be a legal requirement to the algorithmic context, the most we can establish is that a very strange kind of sincerity (one that we would not ordinarily consider) should not be a legal requirement. But that kind of sincerity is neither intuitively desirable nor is it compatible with the action guiding feature of explanations of algorithmic decisions. Meanwhile, Cohen's argument does not diminish the value of folk sincerity, so to speak – i.e., the kind that requires producing motivating reasons. This is the kind of sincerity that is crucial for action guidance.

Now it is true that in the non-algorithmic context, judges may not always be able to produce motivating reasons – as many in the American Legal Realist tradition have argued.[77] We may not even have the requisite luminosity (as philosophers call it) to discern the 'true' reasons motivating our behavior in the first place.[78] This means that in the ordinary (non-algorithmic) context, requiring internalist sincerity would run afoul of the principle (associated first and foremost with Immanuel Kant) that "Ought Implies Can":[79] we cannot require judges to do something that is in fact impossible for them to do. We often tell ourselves rationalizations, convince ourselves of simplistic stories for why we act, and engage in other irrational

---

[76] Once again the example is for illustrative purposes only. We take no position on whether this actually is a justified reason for a particular sentence – we don't seek to offer a worked out theory of the philosophical underpinning of sentencing by American judges. Instead we just offer this as an example. If it is not one you, dear reader, find normatively justified feel free to substitute another. It doesn't matter for our purposes.

[77] JEROME FRANK, LAW AND THE MODERN MIND 108–09, 119–21 (1970); Felix S. Cohen, *Transcendental Nonsense and the Functional Approach*, 35 COLUM L. REV. 809, 849 (1935); O.W. Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 469 (1897). The notion that "law is . . . *only* a matter of what the judges eat for breakfast" has been (with some degree of question) attributed to the realists. *See* Dan Priel, *Law Is What the Judge Had for Breakfast: A Brief History of an Unpalatable Idea*, 68 BUFF. L. REV. 899, 902 (2020) (quoting Ronald Dworkin, *Dissent on Douglas*, N.Y. REV. BOOKS (Feb. 19, 1981), at 4). Yet at least some analysis seems to suggest this might not fully be an exaggeration. Shai Danziger, Jonathan Levav & Liora Avnaim-Pesso, *Extraneous Factors in Judicial Decisions*, 108 PROC. NAT. ACAD. SCI. 6889 (2011). *But see, e.g.*, Keren Weinshall-Margel & John Shapard, *Overlooked Factors in the Analysis of Parole Decisions*, 108 PROC. NAT'L ACAD. SCI. E833 (2011). In particular, the disjuncture between description and normative justification may play some role in judicial difficulties in producing motivating reasons. *See* DAVID HUME, A TREATISE OF HUMAN NATURE 302 (David Fate Norton & Mary J. Norton eds., Oxford Univ. Press 2000) (1739–1740); J.L. MACKIE, ETHICS: INVENTING RIGHT AND WRONG 81 (1977).

[78] *See* TIMOTHY WILLIAMSON, KNOWLEDGE AND ITS LIMITS 107–08 (2000).

[79] *See* IMMANUEL KANT, CRITIQUE OF PURE REASON 637 (Norman Kemp Smith trans., St. Martin's Press, Macmillan, 1929); SAMUEL KAHN, KANT, OUGHT IMPLIES CAN: THE PRINCIPLE OF ALTERNATE POSSIBILITIES, AND HAPPINESS 13–19 (2019).

behavior governed by imperfect heuristics.[80] Indeed, this is why the legal system often appears to practice a kind of externalist sincerity. When an appellate judge reviews a trial court judge's decision, the appellate judge takes the trial judge's opinion at face value and evaluates whether it contains what Cohen would call legitimate normative reasons. It's entirely plausible that the judge is a closet racist and actually decided on the basis of illegitimate motivating reasons. But an appellate judge would look at the stated reasons. And of course the set of stated reasons supporting a decision is not unique. There are many rationales that the trial judge can give to justify a particular choice. What the appellate judge looks for is whether the reasons given are indeed legitimate. But notice that if judges *could* produce motivating reasons, that would be a fantastic trait of their written decisions, which would allow for much more effective appellate review and for self-correction. The main problem in non-algorithmic decision making is that we cannot require judges to do the impossible.

But what is impossible for judges is not impossible for algorithmic decision makers. Hence, in the context of algorithms we can demand sincerity (in the sense of producing motivating reasons) without violating the Ought Implies Can principle.

For algorithms, motivating reasons would roughly correspond to reasons that are not post hoc – i.e., reasons that causally connect the model's features to its prediction. This is also what the Canadian bill on explainability, with which this paper began, calls "principal factors."[81] For example: we might say that someone is admitted to law school if their combined LSAT and GPA score reaches a certain threshold t on a standardized scale. This is a causal explanation *within* the model – any tinkering with either the LSAT or the GPA in a way that produces a score above the threshold will in fact lead to admission given this model. At the risk of being unduly anthropomorphic, we could call this an algorithm's motivating reasons. For interpretable algorithms, such reasons can be produced. When it comes to algorithms, therefore, we have models that can produce the algorithmic analogue to motivating reasons. And hence we should use those models wherever possible. In short, there is an epistemological problem given our own limited cognitive capacities, which makes it difficult or impossible for us to know what our own motivating reasons are. That is why it is reasonable to tolerate internalist insincerity in legal decision making, as Cohen suggests. But this limitation does not exist for AI/ML. Hence, we should not tolerate insincerity in explanations.

D. From Insincerity to Resentment

Just adjacent to the argument we have offered here, pertaining to sincerity and justification, is a still more philosophical question we

[80] *See* Robert A. Prentice, *Behavioral Ethics: Can It Help Lawyers (and Others) Be Their Best Selves?*, 29 NOTRE DAME J.L. ETHICS & PUB. POL'Y 35, 69–72 (2015); Mark Kelman, *Moral Realism and the Heuristics Debate*, 5 J. LEGAL ANALYSIS 339, 347–48 (2013).
[81] Bill C-27, *supra* note 5, § 63(4).

highlight for future work: the relationship of AI/ML, explanation and reactive attitudes. While explanations are particularly useful for action-guiding, where we argued motivating reasons are central, they can also play another more normative role, where motivating reasons can be even more salient: we often want an explanation because we are interested in answering questions about moral responsibility, broadly construed. When we think a mistake has occurred, we want to potentially blame someone. And indeed, when a correct decision is made, we want to praise someone as well. As automated systems become increasingly prevalent, they will also become the subjects of such intentional attitudes.

In particular, as Boric Babic and Zoë Johnson King argue,[82] they can become the subject of second person "reactive" attitudes, as P.F. Strawson calls them,[83] such as blame, praise, contempt, and resentment. As Strawson puts it, "it matters to us, whether the actions of other people . . . reflect attitudes towards us of goodwill, affection, or esteem . . . or contempt, indifference, or malevolence . . . ."[84] This is often described as the quality of will approach to moral responsibility – we care about the quality of will others display toward us.[85] This is where the "fool's gold" nature of explainability algorithms, as we have called it, might be the most fraught.

There are many situations where a decision will naturally invite reactive attitudes among those concerned. For example, suppose that a family member in need of an organ donation is assigned a low rank in an allocation system for one of a limited number of kidneys, where the assignment is based (at least in part) on a prediction of how medically urgent a transplant would be for that patient. Now suppose further that an explainability algorithm is used to generate an explanation as to why this patient was deemed less medically urgent and the explanation identifies age and gender as salient factors in the prediction. Suppose further that the patient or a civil society group finds it inappropriate for such decisions to be based on gender, and wants to condemn this as a "sexist algorithm" (This is not so far-fetched – an organ transplant related algorithm was recently decommissioned for fear it was producing unjustified results that disfavored black patients.[86])

In such cases – cases that invite reactive attitudes – post hoc explanations not only fail to provide the benefits touted of them, they can also further undermine the legitimacy of automated systems by presenting a narrative that seems intentional in a way that it can be

---

[82] Boris Babic & Zoë A. Johnson King, *Algorithmic Resentment* (November 1, 2022) (unpublished manuscript) (on file with author).

[83] *See* STRAWSON, *supra* note 21, at 4-6.

[84] *Id.* at 5.

[85] There is a question here about whether algorithms can be the sort of agents that are subject to intentional attitudes. Can we even evaluate algorithms with respect to their quality of will? Babic and Johnson King argue that algorithms are indeed appropriate moral agents for Strawsonian reactive attitudes. In this project, we adopt their perspective as to algorithmic quality of will. *See* Babic & Johnson King, *supra* note 82.

[86] *See* AST, *Racial Bias in Clinical Tools and Impact on Organ Donation* (2021), *available at* https://www.myast.org/sites/default/files/Racial%20Bias%20in%20Clinical%20Tools%202021.04.21.pdf.

the subject of blame, praise, or resentment, when in fact it is not. Reactive attitudes require motivating reasons. In order to assess someone's quality of will, we need evidence of why they do what they do – we need to know the reasons that make them act the way they do. And indeed, even if they do the right thing but for the wrong reasons, they could still be the subject of blame. For example, imagine an unscrupulous doctor who intentionally recommends unnecessary diagnostic tests to a patient in order to bill as many services as possible. The doctor does not have evidence to believe the patient might be sick, but unbeknownst to him, one of the tests identifies an extremely rare and unsuspected tumor, which is then removed. We can still blame the doctor for his unethical practices, even if in this case he accidentally did the right thing. Meanwhile, if people do the wrong thing for the right reasons, we might still praise them for trying to do the right thing.[87] Sticking with the earlier example, imagine a doctor who has strong and legitimate reasons to believe that a patient has a certain tumor which is malignant. The doctor removes the tumor but it turns out to be benign and the operation, in retrospect, was unnecessary. We can still praise the doctor for acting diligently and responding to evidence and recommending appropriate care as he ought to. In short, the fact that post hoc explanations cannot deliver motivating reasons means that we cannot use them as the grounds for reactive attitudes, because reactive attitudes are particularly attuned to evaluating an agent's quality of will.

There are also cases that involve both action guidance and evaluative attitudes. In such cases, post hoc explanations are most treacherous. And as it turns out, many contexts where we wish to apply algorithmic systems are exactly of this sort. For instance, consider the ubiquitous case of an algorithm which produces financial risk scores of default in order to determine whether an applicant receives a loan or not. First, there is an action guiding component in cases like this: a rejected applicant would like to know what she needs to change in order to receive a loan the next time around. Second, the situation invites reactive attitudes: if we learn that the algorithm is mistakenly denying loans to, say, persons of color or women, at disproportionate rates, we are likely to feel indignant and blame the system. A post hoc explanation in this case will create a blameworthy straw man – it will generate some reasons which are not the actual reasons for the decision.

IV. SHOULD INTERPRETABILITY BE REQUIRED?

In the last two parts we have argued "don't believe the hype" about explainability. Explainable AI/ML fails both to effectively guide action and to produce sincere explanations, and each is a problem for the case made by its scholarly champions as well as the legislators that want to adopt it as a legal requirement.

[87] Johnson King, *supra* note 22, at 427.

By contrast, *interpretable* AI/ML does not produce the same problems – it can guide action and it gives the actual reasons for a result. The natural question then is, should policymakers require *interpretable* AI/ML models? Ultimately, our view is that while such models are desirable from a legal and political perspective, their use should not be mandated across the board. Such a position stands to undermine technological innovation too much. We also consider whether *some* decision making contexts might require interpretable AI/ML models more than others. Here we tentatively conclude that in certain cases, where democratic freedoms or concerns of procedural justice arise, a policy prohibiting opaque models may indeed be appropriate

A. Interpretability, Action Guidance, and Accuracy

While explainable AI/ML models generate a false sense of transparency, as we have argued, using interpretable AI/ML models allows users to understand the motivating reasons behind a decision and it enables them to plan accordingly. In our MIT Law example, a simplistic interpretable model could say something like this: Every student whose sum of their LSAT score divided by 100 and their undergraduate GPA divided by 2 exceeds 3.5 will be admitted. This is a very naive hypothetical rule, but the idea is that it weights an applicant's LSAT and GPA roughly equally, and admits every applicant whose combined score exceeds a certain threshold, in this case 3.5. For example, a student with a 175 LSAT and a 3.8 GPA would have a combined score of 175/100 + 3.8/2 = 3.65. They would be admitted. A student with a perfect GPA (4) would need an LSAT score of at least 150 because they need to satisfy 2 + x/100 > 3.5 where x represents their LSAT score. A student with a perfect LSAT would need a GPA of at least 3.4 because they need to satisfy 1.8 + x/2 > 3.5. This naive rule could be made more intelligent with just a few simple tweaks. For example, MIT Law could add two simple conditions: Every applicant whose combined score exceeds 3.5 will be admitted, provided that also (1) their GPA is at least 3.7 and (2) their LSAT is at least 164.

Such a rule is very effective at guiding behavior. Every applicant knows that if their GPA is under 3.7, or their LSAT score is under 164, it would be a waste of their time and money to apply. They can also compute their threshold score exactly. And if they cannot be admitted this year, they can determine whether it is worthwhile to try and improve their LSAT score, their GPA, or both, in order to apply next year.

So why not always use such simple and transparent rules? Indeed, some authors argue we should.[88] There are several arguments given against interpretable AI/ML models. First, as the number of input variables grows, simple interpretable rules are harder to construct. Imagine if instead of LSAT and GPA, we had 10,000

[88] Rudin, *supra* note 12; Babic et al., *supra* note 13.

observations for each applicant (educational history, extracurricular activities, work experience, their grade in every class taken, etc.). We *could* try to use a threshold rule, but if every applicant had to plug in 10,000 values in order to determine their prospects for admission, the rule would cease to be useful in its action guidance.

Second, it has been argued that there is a fundamental trade-off between accuracy and interpretability, at least in some contexts.[89] While it may be a little bit fanciful to imagine 10,000 values relevant to law school admission, that kind of high dimensionality is the norm when it comes to image recognition or medical forecasting based on genetic history. Indeed, image recognition, natural language processing, and medical forecasting models are often based on millions of input variables. And often, these variables are not intuitively meaningful – for example, they may represent pixel values instead of an applicant's LSAT score. In cases like this, it can be hard to construct an interpretable AI/ML model which is as accurate as the most complex model that is available.

Dziugaite and colleagues articulate a mathematical argument for the existence of a trade-off between interpretability and accuracy.[90] Increasing complexity gives a model more power and more flexibility to approximate highly non-linear feature-label relationships. Indeed, it is well-known that any Borel-measurable function on a finite-dimensional feature space can be approximated arbitrarily accurately by a simple neural network.[91] If we require an AI/ML model to be interpretable, then we allow ourselves to use only a proper subset of the set of all possible models. As a result, it stands to reason that the best model in the proper subset may be worse than the best of all models.[92]

But there are also reasons to doubt the existence of a tradeoff between accuracy and interpretability. Rudin explores many areas where there is no apparent advantage to using black-box models. For example, a simple three-rule model obtained by the Certifiably Optimal Rule Lists (CORELS) algorithm attains the same accuracy as the well-known proprietary COMPAS recidivism model on the Broward County, Florida data.[93] Rudin argues more generally, using a Rashomon-set[94] strategy, that when a classification task is such that many models perform equally well on it, there likely exists one such

---

[89] *E.g.*, Gintare Karolina Dziugaite, Shai Ben-David & Daniel M. Roy, *Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability*, ARXIV (Oct. 28, 2020), https://arxiv.org/abs/2010.13764.

[90] *Id.*

[91] Kurt Hornik, Maxwell Stinchcombe & Halbert White, *Multilayer Feedforward Networks are Universal Approximators*, 2 NEURAL NETWORKS 359 (1989).

[92] *See* Dziugaite et al., *supra* note 83, at 6.

[93] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer & Cynthia Rudin, *Learning Certifiably Optimal Rule Lists for Categorical Data*, 18 J. MACH. LEARNING RSCH., no. 234, 2018, at 1.

[94] A Rashomon set is a set of models that perform similarly well on a task. Lesia Semenova, Cynthia Rudin & Ronald Parr, *On the Existence of Simpler Machine Learning Models*, ARXIV (May 12, 2022), at 2, https://arxiv.org/abs/1908.01755 (presentation at 2022 ACM Conference on Fairness, Accountability, and Transparency). The name comes from Akira Kurosawa's film, Rashomon, in which multiple people describe the murder of a samurai from different perspectives.

AI/ML model that is interpretable.[95] The reason we observe diminished performance is not because interpretable AI/ML models are necessarily worse, but simply because we have not yet identified the most accurate one for that particular task.

Third, *even if* we cannot prove that opaque models necessarily perform better in some case, it can be argued that requiring AI/ML models to be interpretable is a very intrusive policy strategy. In other words, one might argue that a blanket policy prohibiting deep learning models and complex neural networks is antithetical to technological development and innovation. This libertarian argument is particularly compelling when there is still a lot of uncertainty about which models work best and under what types of circumstances. To prohibit manufacturers from using the latest and most exciting algorithms, the argument goes, would simply be too heavy handed.

As a result of the three arguments given above, we do not want to go so far as to suggest that policy makers should require interpretable AI/ML models. What we would like to do instead is to shift the conversation: the default assumption should be that a simple interpretable AI/ML model ought to be used, unless there is some evidence that an opaque model would be more suitable. Currently, in many applications, the default is reversed: the starting point is to apply a deep learning model to just about any task even if the use of such a complex model is not at all motivated.[96]

B. Interpretability and Procedural Justice

Finally, even though we stop short of arguing that interpretable AI/ML models should be required, as opposed to explainable ones, there are some specific contexts where that policy may be wise. For example, imagine a situation where a scarce number of organs is allocated on the basis of an algorithm which determines the most suitable patients for a transplant. For someone who is denied a transplant, especially in a healthcare system that is at least in part public, it would be eminently reasonable for that patient or his physician to inquire on what basis a patient's suitability was determined: to what extent did the algorithm use comorbidities? Age? Smoking status? Predicted longevity? Contribution to society? Income? Marriage status? Ethnicity? Religion? These are not questions we can answer using explainable AI/ML, for the reasons we have argued in this project. But under interpretable AI/ML models, the actual features used will be immediately available. And in a context like allocation of scarce medical resources, where patients may want to

[95] *Id.* at 1.

[96] There is an interesting sociological question about *why* this has become the default, though we can only speculate: A cultural explanation may be that we have a generation of computer scientists who have been trained on cutting edge deep learning and neural networks and when faced with a problem they do what they were trained to do, even if an interpretable model might be just as good. A more economic argument is that the deep learning and neural networks are attractive from an intellectual property and competition perspective: without the underlying data sets, which along with other elements companies try to protect via trade secrecy, they are harder to reverse engineer or even to evaluate.

appeal decisions, and where trust and democratic legitimacy are paramount, it may be prudent to limit ourselves to interpretable AI/ML models.

Indeed, these advantages of interpretable AI/ML may be compelling *even if* there is some accuracy cost and some impediment to technological innovation. To put it sharply, one may prefer a more interpretable AI/ML model for organ allocation that everyone agrees does a less "good" job in deciding which patient gets offered an organ, as compared to the black-box model, but can do so in a way that is more transparent. In some context, we should be willing to trade off a little (or perhaps even a lot) of accuracy for more visibility into the reasons given. Without purporting to map out all the contexts where these trade-offs particularly favor interpretability, we would suggest that a prime example of a case where this is true is the use of algorithms in the criminal justice system.

CONCLUDING REMARKS

In this paper we have suggested that the current enthusiasm among scholars and among policymakers for explainable AI/ML is misplaced. While running a second algorithm to explain a black box seems neat in theory, we have suggested two main reasons why the explanations it offers are not the kinds of explanations worth having: that its explanations fail to be action guiding and that they can be insincere. At the same time, we have not argued that policymakers should adopt a categorical rule rejecting black boxes and requiring interpretable AI/ML. There are some instances where the benefits of a black box might justify its usage, but we do think a strong presumption in favor of interpretable AI/ML that must be overcome before a black box is used might be a good background rule. Importantly, as we have explained, there may be some contexts where even when everyone agrees that an interpretable AI/ML will produce less accurate or otherwise worse results, its benefits as to transparency and procedural justice might justify favoring it.