UNABASHED BIAS: HOW HEALTH-CARE ORGANIZATIONS CAN SIGNIFICANTLY REDUCE BIAS IN THE FACE OF UNACCOUNTABLE AI

RIYAD A. OMAR†

ABSTRACT

In late 2019, researchers reported evidence of "significant racial bias" in a health-care cost-prediction algorithm that impacted tens of millions of Americans. The researchers diagnosed the problem as likely arising from the development of that algorithm. The manufacturer of the algorithm, however, touted the accuracy of the algorithm for its original purpose and called on its customers—health systems and hospitals—to ensure that underserved populations receive effective care. It was difficult to identify the extent and cause of the significant racial bias because the algorithm in question was a "black box," where the developer did not disclose the relationship between the algorithm's inputs and outputs. Moreover, it is unclear whether the manufacturer tested the algorithm for bias or identified any bias risks to its customers.

This situation highlights key problems confronting health-care organizations looking to implement AI-type algorithms to support health-care decision-making. Such algorithms have a propensity to reflect and reinforce societal biases regardless of the manufacturer's intent. The details of their operation are often poorly understood by customers and, at times, even by their developers. This places health-care organizations that implement such algorithms to support clinical decisions in the difficult position of having to diagnose biases that developers may be unwilling to test, disclose, or remedy.

This Article argues that health-care organizations can drive significant improvements in AI accountability by applying already familiar data governance principles to how they implement AI algorithms. The approach is anchored in the recognition of two demonstrated risks associated with the use of AI in decision-making: first, algorithms have a propensity to reflect societal biases; and second, AI developers often market

† J.D, CISP/US, CISP/E. J.D., Stanford Law School; B.A., University of Southern California. Mr. Omar's work seeks to advance best practices in health-care innovation through accountability-by-design approaches. Previously, Mr. Omar was Vice President of Privacy and Security for Clarify Health Solutions, a provider of health-care analytics, and Senior Vice President, General Counsel, Secretary, and Chief Strategy Officer of Practice Fusion, a leading electronic health record technology developer. Mr. Omar thanks Carolyn Welter, Danielle Hartley, Maximillian Parmenter, Harshwinder Kaur, and Emily Maino of the *Denver Law Review* at the University of Denver Sturm College of Law for editing this Article and providing invaluable feedback.

their algorithms for uses that have not been independently evaluated. By adopting a formalized approach to addressing both risks, health-care organizations can significantly reduce the harmful impact of such algorithms, including the perpetuation of racial biases. Additionally, this approach rewards AI developers who proactively address those risks, in particular, the developers who adopt practices that foster transparency in how AI algorithms operate so that defects can be detected and remedied.

TABLE OF CONTENTS

INTRODUCTION: THE INHERENT RISK OF USING BLACK BOX ALGORITHMS

In late 2019, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations* (*Dissecting Bias*)[1] found that a “widely used algorithm, typical of this industry-wide approach and affecting millions of patients” throughout the United States “exhibits significant racial bias.”[2] Similar to previous work on algorithmic bias in the criminal justice system as well as in employment, education, and financial services,[3] *Dissecting Bias* raised the possibility that algorithms used in health-care

[1] Ziad Obermeyer, Brian Powers, Christine Vogeli, & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, SCI., Oct. 25, 2019, at 447.

[2] *Id.*

[3] *See, e.g.*, Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1275–76 (2020); Kimberly A. Houser, *Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making*, 22 STAN. TECH. L. REV. 290, 294 (2019); Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1094 (2018) [hereinafter *Explainable Machines*]; Solon Barocas & Andrew D. Selbst, *Big Data’s Disparate Impact*, 104 CALIF. L. REV. 671, 673–74 (2016) [hereinafter *Big Data*]; Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109, 1112–15 (2017); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1024–25 (2017).

decision-making may also present the risk of perpetuating biases against protected classes. On the day *Dissecting Bias* was published, the New York Department of Financial Services and the New York Department of Health wrote to the developer of the algorithm in question[4] citing *Dissecting Bias* for the proposition that the "flawed algorithm ranked healthier white patients as equally at risk for future health problems . . . as [B]lack patients who suffered from far more chronic illnesses."[5]

Research on algorithmic bias has focused on the inherent propensity of artificial intelligence (AI) and big-data algorithmic decision systems to reflect societal biases. As noted by Professors Anya Prince and Daniel Schwarcz, "[t]his is because increasingly sophisticated AIs will affirmatively 'seek out' proxies for prohibited, but predictive, characteristics within increasingly vast amounts of training data."[6] When it comes to algorithmic decision-making about individuals, being "predictive" is often synonymous with society's biased expectations.[7] In *Big Data's Disparate Impact* (*Big Data*),[8] for example, an algorithm designed to help select medical-school candidates extrapolated that being a white male was indicative of academic success because, not surprisingly, a disproportionate number of historical graduates were white males.[9]

The propensity to reflect societal biases can be thought of as a subset of a more general propensity for AI to extrapolate incorrect conclusions from the data used to train it. In *The Intuitive Appeal of Explainable Machines*, Professors Andrew Selbst and Solon Barocas discussed a seminal study where an AI algorithm concluded that pneumonia patients had a reduced risk of death if they also happened to be asthmatic.[10]

> To anyone with a passing knowledge of asthma and pneumonia, this result was obviously wrong. The model was trained on clinical data from past pneumonia patients, and it turns out that patients who suffer from asthma truly did end up with better outcomes. What the

[4] Letter from Linda A. Lacewell, Superintendent, N.Y. State Dep't of Fin. Servs. & Howard A. Zucker, Comm'r, N.Y. State Dep't of Health to David S. Wichmann, Chief Exec. Officer, UnitedHealth Grp. Inc. (Oct. 25, 2019) [hereinafter New York Letter].

[5] *Id.*

[6] Prince & Schwarcz, *supra* note 3, at 1276.

[7] *See id.* at 1275; *Big Data*, *supra* note 3, at 673–74.

[8] *Big Data*, *supra* note 3.

[9] *Id.* at 682 (citing Stella Lowry & Gordon Macpherson, *A Blot on the Profession*, 296 BRIT. MED. J. 657, 657 (1988)) ("St. George's Hospital, in the United Kingdom, developed a computer program to help sort medical school applicants based on its previous admissions decisions. Those admissions decisions, it turns out, had systematically disfavored racial minorities and women with credentials otherwise equal to other applicants'. In drawing rules from biased prior decisions, St. George's Hospital unknowingly devised an automated process that possessed these very same prejudices. . . . '[T]he program was not introducing new bias but merely reflecting that already in the system.'").

[10] *Explainable Machines*, *supra* note 3, at 1123 (citing Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, & Noémie Elhadad, *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission*, PROC. 21ST ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING at 1721, 1721 (Aug. 2015)).

> model missed was that these patients regularly monitored their breathing, causing them to go to the hospital earlier. Then, once at the hospital, they were considered higher risk, so they received more immediate and focused treatment.[11]

The propensity for AI to extrapolate the wrong conclusions from data is pervasive. *Big Data*[12] describes five mechanisms by which "disproportionate adverse outcomes might occur" whenever developers mine big data to develop algorithms or models: bias in defining the "target variables" or "class labels," bias reflected in training data, bias in data collection, bias in feature selection, and bias in proxies.[13] The authors of *Dissecting Bias*, for example, hypothesize that the root cause of the significant racial bias in the "widely used algorithm" that was the subject of their study may be a problem of "label choice" in the algorithm's development:

> Bias attributable to label choice—the difference between some unobserved optimal prediction and the prediction of an algorithm trained on an observed label—is a useful framework through which to understand bias in algorithms, both in the health sector and further afield. This is because labels are often measured with errors that reflect structural inequalities. Within the health sector, using mortality or readmission rates to measure hospital performance penalizes those serving poor or non-[w]hite populations.[14]

The algorithm's manufacturer, however, defended the efficacy of its algorithm for its designed purpose. The algorithm, the manufacturer argued, is "highly predictive of [future medical] cost, which is what it was designed to do."[15] Moreover, as *Dissecting Bias* notes, the algorithm does not appear to exhibit significant racial bias when the algorithm is used for that specific purpose.[16]

This raises two key questions: Was the algorithm being used for purposes beyond "what it was designed to do," and if so, did such uses give rise to the observed significant racial bias? *Dissecting Bias* suggests the answer to both questions is yes.

[11] *Id.*

[12] *Big Data*, *supra* note 3.

[13] See *id.* at 677–91, for a more in-depth discussion of the five mechanisms.

[14] Obermeyer et al., *supra* note 1, at 452–53.

[15] Christopher Snowbeck, *Regulators Probe Racial Bias with UnitedHealth Algorithm*, STAR TRIB. (Oct. 28, 2019, 6:59 PM), https://www.startribune.com/regulators-probe-racial-bias-with-unitedhealth-algorithm/563997722/.

[16] Obermeyer et al., *supra* note 1, at 450–51 ("As a first check . . . we calculate the distribution of realized costs *C* versus predicted costs *R*. By this metric, *one could call the algorithm unbiased*. . . . [A]t every level of algorithm-predicted risk, Blacks and [w]hites have (roughly) the same costs the following year. . . . Conditional on risk score, predictions do not favor [w]hites or Blacks. . . .") (emphasis added).

The hospital studied in *Dissecting Bias*, for example, did not use the algorithm's risk scores to predict its patients' medical expenditures, which is what the algorithm was designed to do. Rather, the hospital used the risk scores to identify which of its patients suffered from multiple chronic medical conditions.[17] Patients identified as high risk would receive "greater attention from trained providers, to help ensure that care is well coordinated,"[18] including "teams of dedicated nurses, extra primary care appointment slots, and other scarce resources[,]" that "are widely considered effective at improving [health-care] outcomes."[19] The risk scores were being used to support medical decisions that impacted which patients would receive medical care.

This expanded use of the algorithm appeared to be endorsed by the manufacturer. In a marketing brochure quoted by *Dissecting Bias*, the manufacturer touted the algorithm's ability to "determine which individuals are in need of specialized intervention programs and which intervention programs" and to "flag individuals for intervention before their health becomes catastrophic." [20] This expanded use, however, also appeared to give rise to the observed significant racial bias. *Dissecting Bias* found that, at a given risk score assigned by the algorithm, "Black patients are considerably sicker than [w]hite patients."[21] Black patients receiving the same "risk score" had "26.3% more chronic illnesses than [w]hites."[22] Because Black patients' health risks were underestimated,[23] there was a significant risk that Black patients who would otherwise qualify for much needed medical services were being overlooked.[24] Using the algorithm's risk scores as a measure of severity, for example, only 17.7% of patients in the highest risk category were Black; if, however, the hospital used the study's independently developed measure of comorbidity, that percentage would rise to 46.5%.[25] Customers using

[17] Obermeyer et al., *supra* note 1, at 447–48.

[18] *Id.* at 447.

[19] *Id.*

[20] *Impact Pro for Population Health Management (2014)*, OPTUM (on file with author); Ziad Obermeyer, Brian Powers, Christine Vogeli, & Sendhil Mullainathan, *Supplementary Materials for Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, SCI. 1, 3 (2019), https://science.sciencemag.org/content/sci/suppl/2019/10/23/366.6464.447.DC1/aax2342_Obermeyer_SM.pdf [hereinafter *Supplementary Materials*] ("The algorithm's stated goal (from promotional materials) is to predict *which individuals are in need of specialized intervention programs and which intervention programs have the most impact on the quality of individuals' health*. These scores, which are meant to *flag individuals for intervention before their health becomes catastrophic*, are a key part of the decision to enroll a patient in the care management program . . . .") (quoting the Optum *Impact Pro* brochure).

[21] Obermeyer et al., *supra* note 1, at 447.

[22] *Id.* at 448.

[23] *See* New York Letter, *supra* note 4 ("[The] flawed algorithm ranked healthier white patients as equally at risk for future health problems . . . as [B]lack patients who suffered from far more chronic illnesses.").

[24] *Id.* ("[The] algorithm appears to inherently prioritize white patients who have had greater access to healthcare than [B]lack patients.").

[25] Obermeyer et al., *supra* note 1, at 449 ("[A]t α = 97th percentile, among those auto-identified for the program, the fraction of [B]lack patients would rise from 17.7 to 46.5%.").

these risk scores to enroll their patients in specialized care management programs, therefore, would be enrolling significantly fewer Black patients than warranted by the patients' actual medical conditions.

The algorithm's manufacturer appeared to be aware that the use of its risk scores to identify a patient's medical conditions could result in "gaps" adversely impacting certain communities.[26] These gaps, the manufacturer noted, are "caused by social determinants of care and other socio-economic factors."[27]

This raises the question of whose responsibility it is to identify these gaps and mitigate their harmful effects. According to the manufacturer, the responsibility falls solely to its customers. "These gaps," the manufacturer states, "can . . . be addressed by the health systems and doctors to ensure people, especially in underserved populations, get effective, individualized care."[28] The manufacturer's current marketing materials further emphasize that customers are responsible for the proper implementation of the manufacturer's algorithms, noting that the algorithms are there to support "clinician-driven identification and stratification" of health conditions, where the algorithm "help[s] you flag individuals."[29]

According to *Dissecting Bias*, however, it is not easy for customers to investigate the extent to which the algorithm exhibits bias.[30] This is because the algorithm in question is a "black box" where the manufacturer does not disclose how its algorithm operates and customers cannot "observe how the[] raw data are combined to form the specific variables used for prediction."[31] Consequently:

> Empirical investigations of algorithmic bias . . . have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work 'from the outside,' often with great ingenuity, and resort to clever work-arounds such as audit studies.[32]

[26] Snowbeck, *supra* note 15 ("[T]he cost model within Impact Pro was highly predictive of cost, which is what it was designed to do.") (quoting UnitedHealth Group statement).

[27] *Id.* (quoting UnitedHealth Group statement).

[28] *Id.* (quoting UnitedHealth Group statement).

[29] *Impact Pro: Better Understand Individual and Population Health Needs*, OPTUM, https://www.optum.com/content/dam/optum3/optum/en/resources/sell-sheet/impact-pro-sell-sheet.pdf (last visited June 16, 2021)

[30] Obermeyer et al., *supra* note 1, at 447.

[31] *Supplementary Materials*, *supra* note 20, at 3 ("[W]e observe the raw insurance claims data that form the totality of inputs to the predictive algorithm (though we do not observe how these raw data are combined to form the specific variables used for prediction.").

[32] Obermeyer et al., *supra* note 1, at 447.

With a black-box algorithm, the customer does not know "how factors are weighed or how risk scores are determined."[33] All the customer gets is an output such as a risk score.[34] As a result, it is often difficult, if not impossible, for a user to understand when an error has occurred.

If, for example, a black-box algorithm assigns a low score to a given patient based on an undisclosed formula, a customer has no straightforward way of detecting whether or not the low score is correct. This inability to detect defects, in turn, allows algorithms to perpetuate errors, including biases, indefinitely without detection. Indeed, the algorithm's biased outputs themselves can create the circumstances that validate their own faulty predictions. An erroneously assigned low credit score, for example, can result in an individual losing the ability to obtain credit; this inability to obtain credit, in turn, ends up "confirming" the algorithm's original erroneous score because the individual was never able to establish their creditworthiness. The individual became a credit risk because the algorithm said they were.

The practical inability to recognize, address, and correct defects in the algorithms used to make decisions about individuals risks turning AI's potential benefits on their head. Rather than creating algorithms that inform reliable, evidence-based decisions that improve the lives of individuals, opaque business practices create conditions in which algorithms are developed and implemented in ways that accelerate societal biases.[35] This opacity, in turn, can result in defective predictions being regarded as "confirmations," further ingraining, rather than correcting, the initial defects.[36] This phenomenon is discussed in *Runaway Feedback Loops in Predictive Policing,*[37] which examines a "predictive policing" algorithm used by police to determine which neighborhoods to surveil:

> Predictive policing is increasingly employed to determine where to send police, who to target for surveillance, and even who may be a future crime victim. . . . Once police are deployed based on these predictions, data from observations in the neighborhood is then used to further update the model. . . . Since such discovered incidents only occur in neighborhoods that police have been sent to *by the predic-*

[33] State v. Loomis, 881 N.W.2d 749, 763–64 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017) ("[T]he proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined.").

[34] *Id.* at 754.

[35] Lee Rainie & Janna Anderson, *Code-Dependent: Pros and Cons of the Algorithm Age*, PEW RSCH. CTR. (Feb. 8, 2017), https://www.pewresearch.org/internet/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/.

[36] *See, e.g.*, Kristian Lum & William Isaac, *To Predict and Serve?*, 13 SIGNIFICANCE 14, 16 (2016) (discussing how feedback loops based on biases affect over policing in certain areas for crime).

[37] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, & Suresh Venkatasubramanian, *Runaway Feedback Loops in Predictive Policing*, 81 PROC. MACH. LEARNING RSCH. (2018).

> *tive policing algorithm itself*, there is the potential for this sampling bias to be compounded, causing a runaway feedback loop.[38]

The literature on algorithmic bias proposes various solutions. The authors of *Dissecting Bias*, for example, offer quality improvement recommendations in the algorithm's development:

> [W]e must change the data we feed the algorithm—specifically, the labels we give it. Producing new labels requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment.[39]

The problem with this solution, however, is that it requires the developer to accept an obligation that the developer expressly disclaims. According to the algorithm's developer, any gaps in the algorithm's predictions should be "addressed by the health systems and doctors to ensure people, especially in underserved populations, get effective, individualized care."[40] It is the customers, after all, who are choosing to automate their decision-making with a potentially biased algorithm. Accordingly, it is the responsibility of those customers—in this case health systems and doctors—to ensure that they implement such algorithms in a manner that does not adversely impact their patients.

This raises the question of whether it is possible to responsibly implement black-box algorithms given their inherent risks. Did the hospital described in *Dissecting Bias*, for example, address the gaps in the algorithm's risk scores to ensure that patients in underserved populations were not adversely impacted by the algorithm's observed bias? If so, how did the hospital accomplish this in light of the limited information it had about how the algorithm worked?

Dissecting Bias does not discuss whether the hospital conducted a formal risk assessment of the inherent risks in deploying the algorithm. The study, however, provides details of the hospital's implementation[41]—a number of which suggest that the hospital sought to address the algorithm's gaps in an effort to ensure that its patients obtained individualized care.[42] Moreover, it appears that its implementation reduced, at least in part, the significant racial bias reported in *Dissecting Bias*. As such, the hospital's implementation can serve as a case study on how organizations can responsibly implement black-box algorithms in situations where they have incomplete information about how the algorithm works, the extent of the algorithm's accuracy, and the prevalence of algorithmic bias.

[38] *Id.* at 2.
[39] Obermeyer et al., *supra* note 1, at 453.
[40] Snowbeck, *supra* note 15.
[41] *See* discussion *infra* Section I.C.
[42] *Id.*

As will be discussed in Part I, health-care organizations may be in a better position than most to address the inherent risks of algorithmic bias for multiple reasons. First, they often have domain expertise in assessing the strengths and weaknesses of algorithms and devices used in health-care delivery.[43] Second, they often have greater awareness of the benefits of the scientific method's requirements for transparency and the substantial risks of departing from that transparency.[44] Finally, most health-care organizations have adopted risk-management processes governing the utilization of patient-health information—processes that can readily be adapted by such organizations to address the inherent risks of the data generated from black-box algorithms.[45] Something akin to these factors appears to have been at play in how the hospital described in *Dissecting Bias* sought to assess and address the inherent risks of the black-box algorithm.[46]

The Conclusion then examines the benefits of formalizing such risk-management practices by entities that delegate their decision-making to algorithms. A formalized risk assessment, for example, would require such entities to understand the predictive accuracy of the algorithms they implement for the specific questions that are being delegated. For example, how many false positives and false negatives is the algorithm likely to produce? Are those false positives and false negatives clustered in a manner that disparately impacts specific communities? Thereafter, a formalized risk mitigation plan would allow the entity to insulate its patients from the adverse consequences of the algorithm's predicted error rate. Customers would address the gaps that AI developers currently externalize to their customers. Over time, certain developers would likely seize the opportunity to compete on transparency and independently validated accuracy. This, in turn, would drive significant, independently confirmed advances in algorithmic decision-making.

I. Responsible Use of Black-Box Algorithms

A. Health-Care Organizations, Scientific Method and Information Risk Management

Health-care organizations are in a better position than most to assess and mitigate the risks of algorithmic bias. A hospital's "customers" are patients treated in accordance with rigorous, evidence-based interventions.[47] Health-care professionals are familiar with the inherent challeng-

[43] *See* discussion *infra* Section I.A.

[44] *Id.*

[45] *Id.*

[46] *See* discussion *infra* Section I.C.

[47] *See* Marita G. Titler, *The Evidence for Evidence-Based Practice Implementation*, *in* Patient Safety and Quality: An Evidence-Based Handbook for Nurses (R.G. Hughes ed., 2008) ("Evidence-based practice . . . is the conscientious and judicious use of current best evidence in conjunction with clinical expertise and patient values to guide health care decisions.").

es in assessing the medical conditions of a patient in the absence of an express diagnosis.[48] Unlike a financial services company or a governmental body that may uncritically accept a risk score without regard to the magnitude of an algorithm's error rate, health-care organizations are familiar with the limitations of various medical testing capabilities and have developed protocols to detect and address systems that produce false positives or false negatives.[49]

Health-care organizations also routinely deploy safeguards aimed at detecting and remedying operational defects.[50] Large health-care organizations, for example, often deploy quality-management systems where operational requirements are documented and tested through quality-assurance processes.[51] Many organizations are also required to have mature data protection and compliance operations[52] that are well positioned to review algorithms to ensure that their specifications comply with applicable laws. Hospitals and health systems frequently have medical officers who oversee clinical operations and are versed in a wide range of issues surrounding health-care delivery, including the fact that minority groups face significantly greater barriers to accessing medical care.[53] Hospitals and health systems also have research and data science exper-

[48] Viraj Bhise, Suja S. Rajan, Dean F. Sittig, Robert O. Morgan, Pooja Chaudhary, & Hardeep Singh, *Defining and Measuring Diagnostic Uncertainty in Medicine: A Systematic Review*, 33 J. GEN. INTERNAL MED. 103, 103–04 (2018).

[49] *See, e.g.*, *Avoiding Inappropriate Clinical Decisions Based on False-Positive Human Chorionic Gonadotropin Test Results*, AM. COLL. OF OBSTETRICIANS & GYNECOLOGISTS (Nov. 2002), https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2002/11/avoiding-inappropriate-clinical-decisions-based-on-false-positive-human-chorionic-gonadotropin-test-results (offering recommendations for managing false-positive results); *Tuberculosis (TB) False-Positive Investigation Toolkit*, CTRS FOR DISEASE CONTROL & PREVENTION, https://www.cdc.gov/tb/publications/guidestoolkits/false_positive/False-Positive.htm (last visited Apr. 18, 2021) (providing resources for detecting and addressing false-positive tests for tuberculosis).

[50] Anubha Aggarwal, Himanshu Aeran, & Manu Rathee, *Quality Management in Healthcare: The Pivotal Desideratum*, 9 J. ORAL BIOLOGY & CRANIOFACIAL RSCH 180, 180 (2019).

[51] *See, e.g.*, Ian R. Lazarus & M. Wes Chapman, *"ISO-Style" Healthcare: Designed to Keep Patients, Practitioners and Management Safe*, BECKER'S HOSP. REV. (Sep. 26, 2013), https://www.beckershospitalreview.com/hospital-management-administration/iso-style-healthcare-designed-to-keep-patients-practitioners-and-management-safe.html (noting that over 5,000 hospitals and 10,000 other institutions are accredited as implementing a quality management system); 45 C.F.R. § 170.315(g)(4) (2020) (requiring that all capabilities of electronic health record technology that are certified for use in federal health-care quality improvement programs were developed, tested, implemented, and maintained in conformance with a quality management system (QMS) established by the federal government or mapped to such QMS).

[52] *See, e.g.*, 45 C.F.R § 164.530(c)(1) (2009) ("A covered entity [health care provider, health plan or health care clearinghouse] must have in place appropriate administrative, technical, and physical safeguards to protect the privacy of protected health information."); 45 C.F.R. § 164.308(a)(4)(i) (2013) ("[Healthcare organizations must] implement policies and procedures for authorizing access to electronic protected health information that are consistent with the applicable requirements of [HIPAA's Privacy Rule].").

[53] *See generally 100 Hospital and Health System CMOs to Know*, BECKER'S HOSP. REV. (Feb. 19, 2020), https://www.beckershospitalreview.com/lists/100-hospital-and-health-system-cmos-to-know-2020.html (listing hospital CMOs who oversee clinical operations in their hospitals).

tise that can be utilized to validate how algorithms function before and after deployment in clinical operations.[54]

In addition, health-care organizations are familiar with managing information risks that could impact health-care delivery. For example, the Health Insurance Portability and Accountability Act's (HIPAA) security rule[55] requires health-care organizations[56] to "[e]nsure the confidentiality, integrity, and availability of all [patient information that it] creates, receives, maintains, or transmits."[57] In furtherance of this obligation, health-care organizations must conduct "an *accurate* and *thorough* assessment of the *potential risks* and *vulnerabilities* to the . . . integrity" of the health information they hold.[58] Under HIPAA's security rule, the term *integrity* usually pertains to assurances that patient information is not "altered or destroyed in an unauthorized manner."[59] However, this preexisting understanding could readily be adapted to cover the reliability of the outputs of decision-making algorithms. This, in turn, would bring the implementation of decision-making algorithms within a familiar risk-management process that already safeguards the integrity of patient health-care information.

B. An Accurate and Thorough Assessment of Black-Box Algorithms

1. Observability and Algorithmic Accuracy

The miracles of AI have received sizable attention. One of them is AlphaZero, which mastered the game of chess in nine hours.[60] Successes like this one are due, in large part, to the significant advances in computing power. "Over the course of nine hours, [the AlphaZero] played forty-four million games against itself on a massive cluster of specialized

[54] *See, e.g.*, *Research*, CLEVELAND CLINIC, https://my.clevelandclinic.org/research (last visited Apr. 18, 2021) ("Since its founding in 1921, research has been an integral part of Cleveland Clinic's mission. . . . Our researchers are leaders in growing fields transforming the way medicine is delivered, including precision medicine, genomics, population health and immuno-oncology."); *Research & Discovery at UChicago Medicine*, UNIV. CHI. MED., https://www.uchicagomedicine.org/research (last visited Apr. 18, 2021) ("At the University of Chicago Medicine, we translate fundamental scientific discoveries into better care for patients. . . . We perform more clinical trials than any other hospital in Illinois.").

[55] 45 C.F.R. § 160 (2013); 45 C.F.R. § 164(A) (2013); 45 C.F.R. § 164(C) (2013).

[56] 45 C.F.R. § 160.310 (2013) (HIPAA's provisions apply to health-care organizations that qualify as "covered entities" and "business associates.").

[57] 45 C.F.R. § 164.306(a)(1) (2013) ("Covered entities and business associates must . . . [e]nsure the confidentiality, integrity, and availability of all electronic protected health information the covered entity or business associate creates, receives, maintains, or transmits.").

[58] 45 C.F.R. § 164.308(a)(1)(ii)(A) (2013) (emphasis added).

[59] *See* 45 C.F.R. § 164.304 (2013) ("*Integrity* means the property that data or information have not been altered or destroyed in an unauthorized manner.").

[60] James Somers, *How the Artificial-Intelligence Program AlphaZero Mastered its Games*, THE NEW YORKER (Dec. 28, 2018), https://www.newyorker.com/science/elements/how-the-artificial-intelligence-program-alphazero-mastered-its-games.

Google hardware."[61] No human could play forty-four million games in a lifetime, much less in nine hours.

The power of AI, however, does not inevitably create accurate or useful algorithms. In 2016, for example, Microsoft released an AI-driven "chatbot" named "Tay" that interacted with Twitter users.[62] Microsoft unveiled the chatbot as an experiment in "conversational understanding," with the bot learning from its interactions with the Twitter community.[63] Tay began the day posting affable greetings, such as "can i just say that im stoked to meet u? humans are super cool."[64] Soon, however, Tay absorbed the vernacular of Twitter users.[65] Within a few hours of Tay's introduction to the world, the AI-driven chatbot expressed antipathy for humanity ("I just hate everybody"), as well as outright racism, sexism, and anti-Semitism.[66] AI's computational power had created a bigot bot as quickly as AlphaZero became a chess master.

AI is an efficient means of creating complex algorithms, both good and bad. Whether algorithms are socially beneficial depends critically on how they are developed and how transparently their outputs are evaluated. In the game of chess, for example, a "win," "loss," and "draw" can be unambiguously defined. This, in turn, allows AI such as AlphaZero to be trained to seek wins and reinforce the AI whenever it achieves a win. That clarity about what counts as "correct" allows the AI to extrapolate the correct model when it plays its forty-four million games in nine hours.

If, on the other hand, an algorithm's success criteria are defined imprecisely or those criteria incorporate unarticulated assumptions, then AI is an effective means for creating algorithms that exploit such unarticulated assumptions. For example, how do you define what it means to be a successful chatbot? If a developer defines an algorithm's success as "creating Twitter posts that result in the largest number of responses," the AI's reinforcement will generate racist posts whenever such posts result in the greatest number of responses.[67] Microsoft never intended for Tay to be a bigot, and neither Microsoft nor Twitter intended to allow a raving bigot to run loose on Twitter.[68] But Tay's programming resulted in its conclusion that making racist, sexist, and anti-Semitic statements is equivalent to being a successful chatbot.

[61] *Id.*

[62] *See* James Vincent, *Twitter Taught Microsoft's AI Chatbot to Be a Racist [******] in Less than a Day*," THE VERGE (Mar. 24, 2016, 6:43 AM), https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[63] *Id.* ("The more you chat with Tay, said Microsoft, the smarter it gets, learning to engage people through 'casual and playful conversation.'").

[64] *Id.*

[65] *Id.*

[66] *Id.*

[67] *See supra* notes 35–36 and accompanying text.

[68] Vincent, *supra* note 62.

Defining what it means to be a successful chatbot on Twitter is difficult for two reasons. First, there are many implicit assumptions that underly any definition of "good conversationalist," many of which are normative (e.g., "don't be rude"). Second, unlike chess, there may be no consensus about what those assumptions are (e.g., some Twitter users may believe rudeness is acceptable). A critical mass of individuals who care about chess agree on what "winning" means. This is likely not the case with the definition of a "good conversationalist." For some, in fact, generating a lot of clicks or responses, regardless of the substance, may qualify as success.[69]

Even when there is no consensus on the definition of success, AI can be effective when there is consensus on how algorithmic outputs are judged.[70] Regardless of how Tay was programmed, most human observers quickly recognized that Tay's outputs were defective.[71] Thanks to this consensus, and the fact that Twitter users could directly observe Tay's outputs and their impact, Microsoft took Tay offline to make adjustments.[72]

What would have happened, however, if Tay's outputs were not directly observable? What if Tay was instead programed to promote the most "provocative" posts? In this scenario, rather than making racist posts on Twitter, Tay assigned "scores" to other users' posts and instead of users being able to see Tay's scores, Tay's scores were used to "retweet" and promote the "highest scoring" of the other users' provocative posts. If Twitter users did not even know Tay was operating in the background, would they be able to detect whether or not Tay was specifically giving high scores to racist, sexist, and anti-Semitic content? And if users

[69] *See, e.g.*, Naomi Craker & Evita March, *The Dark Side of Facebook®: The Dark Tetrad, Negative Social Potency, and Trolling Behaviours*, 102 PERSONALITY & INDIVIDUAL DIFFERENCES 79, 84 (July 2, 2016) (reporting that negative social rewards attained by online trolls were better predictors of online harassment than the negative personality traits of the harassers).

[70] *See generally NIST Asks A.I. to Explain Itself*, NAT'L INST. OF STANDARDS & TECH. (Aug. 18, 2020), https://www.nist/gov/news-events/news/2020/08/nist-asks-ai-explain-itself (discussing the effectiveness of explainable AI and four principles to judge whether AI is explainable); Reuben Binns & Valeria Gallo, *Accuracy of AI System Outputs and Performance Measures*, INFO. COMM'R'S OFF. (May 2, 2019), https://ico.org.uk/about-the-ico/news-and-events/ai-blog-accuracy-of-ai-system-outputs-and-performance-measures/ (discussing using accuracy of data outputs as a means of measuring AI performance).

[71] *See, e.g.*, Vincent, *supra* note 62 (discussing Tay beginning to repeat misogynistic and racist sentiments); Elle Hunt, *Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter*, THE GUARDIAN (Mar. 24, 2016, 2:41 PM), https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (noting how Tay's Twitter conversations changed to racist and other inflammatory statements); Caroline Sinders, *Microsoft's Tay is an Example of Bad Design*, MEDIUM (Mar. 24, 2016), https://medium.com/@carolinesinders/microsoft-s-tay-is-an-example-of-bad-design-d4e65bb2569f (discussing Tay's bad design in allowing for users to say whatever they want to Tay and allowing Tay to repeat those sentiments).

[72] *See* Vincent, *supra* note 62 (quoting Microsoft's response) ("The AI chatbot Tay is a machine learning project. . . . [S]ome of its responses are inappropriate and indicative of the types of interactions some people are having with it. We're making some adjustments to Tay.").

are unable to detect it, how long would it take for Twitter to recognize that Tay was promoting racist, sexist, and anti-Semitic content?

The Correctional Offender Management and Profiling Alternative Sanctions (COMPAS) is a black-box algorithm used throughout the U.S. criminal justice system.[73] COMPAS utilizes 137 variables[74] to generate risk scores for a defendant or inmate corresponding to pretrial recidivism risk, general recidivism risk, and violent recidivism risk,[75] each on a scale of 1–10.[76] COMPAS's manufacturer does not disclose how COMPAS's risk scores are determined.[77] If, therefore, COMPAS assigns a high recidivism risk score to an eighteen-year-old Black girl arrested for petty theft, it would be impossible for a judge to inspect the risk score, compare it to the 137 variables, and determine whether that risk score is correct, a false positive, or reflective of algorithmic bias. Indeed, judges may not be aware that COMPAS risk scores are capable of producing inaccurate predictions about individual defendants.[78]

This is notably different from interpretable algorithms, where the relationship between inputs and outputs is publicly disclosed. The Framingham Risk Score (FRS), for example, estimates the risk of a patient developing cardiovascular disease (CVD) within ten years based on a patient's age, gender, cholesterol, systolic blood pressure, and HDL-C count as well as whether the patient is a smoker or diabetic.[79] If a software solution computed an FRS of twenty-three for an eighteen-year-old non-smoker and non-diabetic patient who has normal cholesterol, blood pressure, and HDL-C count, the doctor would immediately recognize that an error has occurred. Unlike COMPAS risk scores, an FRS is inter-

[73] Cynthia Rudin, Caroline Wang, & Beau Coker, *The Age of Secrecy and Unfairness in Recidivism Prediction*, HARV. DATA SCI. REV. 1, 2 (Mar. 31, 2020) https://hdsr.mitpress.mit.edu/pub/7z10o269/release/3 ("COMPAS is used throughout the criminal justice system in the U.S., and its predictions have serious consequences in the lives of many people.").

[74] Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, SCI. ADVANCES: RSCH METHODS (Jan. 17, 2018), https://advances.sciencemag.org/content/4/1/eaao5580 (noting that the recidivism risk score algorithm, COMPAS, utilizes 137 input features).

[75] State v. Loomis, 881 N.W.2d 749, 754 (Wis. 2016) ("A COMPAS report consists of a risk assessment designed to predict recidivism. . . . The risk assessment portion of COMPAS generates risk scores displayed in the form of a bar chart, with three bars that represent pretrial recidivism risk, general recidivism risk, and violent recidivism risk.").

[76] *Id.* ("Each bar indicates a defendant's level of risk on a scale of one to ten.").

[77] *Id.* at 761 ("Northpointe, Inc., the developer of COMPAS, considers COMPAS a proprietary instrument and a trade secret. . . . [I]t does not disclose how the risk scores are determined or how the factors are weighed.").

[78] *See, e.g.*, Julia Angwin, Jeff Larson, Surya Mattu, & Lauren Kirchner, *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing ("Judge Babler reduced Zilly's sentence, from two years in prison to 18 months. 'Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months. . . .'").

[79] *See Framingham Risk Score (FRS), Estimation of 10-year Cardiovascular Disease (CVD) Risk*, CANADIAN CARDIOVASCULAR SOC'Y (2017), https://ccs.ca/app/uploads/2020/12/FRS_eng_2017_fnl_greyscale.pdf.

pretable by professionals because users can independently confirm that an FRS is accurately derived from their inputs.[80]

This lack of transparency makes it difficult, if not impossible, for those utilizing black-box algorithms to detect when an algorithm's predictions are erroneous. As noted in *Dissecting Bias*, the practice of black boxing algorithms likewise hinders empirical investigations of algorithmic bias.[81] Unlike Microsoft's chatbot, Tay, where the defective outputs were immediately detected, the errors in black-box algorithms can persist indefinitely without detection by customers and the individuals adversely impacted by the algorithm's predictions. This, in turn, deprives black-box algorithms of the accurate data needed to improve their accuracy.[82] Black-box algorithms, therefore, have a propensity for ambiguous error rates.[83] When the National Center for State Courts (NCSC) surveyed multiple independent assessments of a criminal recidivism algorithm, the NCSC reported large discrepancies in the estimates of the algorithm's accuracy.[84] For customers looking to implement algorithms in their decision-making, black-box algorithms are inherently riskier and more difficult to independently assess.

2. Assessing the Accuracy of Black-Box Algorithms

The manufacturer of the black-box algorithm analyzed in *Dissecting Bias* described the algorithm as "highly predictive of [medical] cost, which is what it was designed to do."[85] As assuring as this statement may initially appear, it presents two problems for a hospital seeking to use the algorithm to identify patients with multiple chronic conditions. First, the phrase "highly predictive" lacks mathematical precision. A hospital can-

[80] *Compare* Matt Henry, *Risk Assessment: Explained*, THE APPEAL (Dec. 14, 2019), https://theappeal.org/the-lab/explainers/risk-assessment-explained/#the-problems-of-interpretation (discussing the problems of interpreting algorithms like COMPAS) *with* Peter M. Brindle, Alex McConnachie, Mark N. Upton, Carole L. Hart, George Davey Smith, & Graham C.M. Watt, *The Accuracy of the Framingham Risk-Score in Different Socioeconomic Groups: a Prospective Study*, 55 BRITISH J. GEN. PRAC. 838 (2005) (discussing specific findings of issues within the predictability of FRSs).

[81] Obermeyer et al., *supra* note 1, at 447.

[82] *See, e.g.*, JUDITH HURWITZ & DANIEL KIRSCH, MACHINE LEARNING FOR DUMMIES 8 (2006) ("Providing accurate machine learning models requires that the source data be accurate and meaningful."). "If you create a machine learning application based on inaccurate data, the application will fail." *Id.* at 37. "One of the great benefits of machine learning is the fact that it requires a constant ingestion of new data in order to be able to make accurate predictions." *Id.* at 53.

[83] *See* Jungwoo Ha, Christopher J. Rossbach, Jason V. Davis, Indrajit Roy, Hany E. Ramadan, Donald E. Porter, David L. Chen, & Emmet Witchel, *Improved Error Reporting for Software that Uses Black-Box Components*, DEP'T OF COMPUT. SCIS. UNIV. AUSTIN TEX. (2007) (explaining ambiguous error behavior in black-box algorithms).

[84] Pamela M. Casey, Jennifer K. Elek, Roger K. Warren, Fred Cheesman, Matt Kleiman, & Brian Ostrom, *Offender Risk & Needs Assessment Instruments: A Primer for Courts*, NAT'L CTR FOR STATE CTS. A-23 (2014) (summarizing the results of COMPAS validation studies that observed AUC results of 0.50 to 0.73).

[85] Snowbeck, *supra* note 15.

not plug the words "highly predictive" into the formula for sensitivity[86] or specificity[87] in order to mathematically calculate how many false positives or false negatives the algorithm is likely to produce. Second, the manufacturer's statement only touts the algorithm's ability to predict medical costs,[88] not the patient's underlying medical conditions.

The conventional approach to measure the accuracy of an algorithm's predictions is to measure its sensitivity and specificity.[89] An instrument's *sensitivity* is also known as its true positive rate because it measures the instrument's ability to identify individuals who have the relevant medical condition.[90] Researchers measure sensitivity by "the percentage of individuals that are correctly identified as being among [the target group]."[91] If, for example, one hundred patients have multiple chronic conditions, an algorithm with a sensitivity of 95% would correctly predict that ninety-five of those individuals have multiple chronic conditions. An instrument's s*pecificity* (or true negative rate), on the other hand, measures the ability to correctly identify when an individual does not belong in the group.[92] Analysts measure specificity by "the percentage of individuals that are correctly identified as *not* being in [that target group]."[93] If one hundred patients do not have multiple chronic conditions, an algorithm with a specificity of 90% would correctly identify ninety of those patients.

An algorithm's outputs are not always dichotomous (e.g., simple yes/no or positive/negative predictions). In the COMPAS algorithm, for example, pretrial recidivism risk, general recidivism risk, and violent recidivism risk[94] are represented by scores on a scale of 1–10.[95] The

[86] *See, e.g.*, Cynthia Rudin & Berk Ustun, *Optimized Scoring Systems: Towards Trust in Machine Learning for Healthcare and Criminal Justice*, WAGNER PRIZE J. at 12 (2018) (unpublished manuscript), https://users.cs.duke.edu/~cynthia/docs/WagnerPrizeJournal.pdf ("The true positive rate (TPR) is the fraction of positive test observations predicted to be positive. Sensitivity is also the true positive rate.").

[87] *Id.* ("Specificity is the true negative rate, the fraction of negative test observations predicted to be negative.").

[88] Snowbeck, *supra* note 15 ("[T]he cost model within Impact Pro was highly predictive of cost, which is what it was designed to do.").

[89] Karimollah Hajian-Tilaki, *Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation*, 4(2) CASPIAN J. INTERN. MED. 627, 627 (2013) ("In diagnostic test with dichotomous outcome (positive/negative test results), the conventional approach of diagnostic test evaluation uses sensitivity and specificity as measures of accuracy of test. . . .").

[90] *See, e.g.*, Rudin & Ustun, *supra* note 86, at 12 ("The true positive rate (TPR) is the fraction of positive test observations predicted to be positive. Sensitivity is also the true positive rate.").

[91] Geof Hileman & Spenser Steele, *Society of Actuaries, Accuracy of Claims-Based Risk Scoring Models*, SOC'Y OF ACTUARIES 44 (2016) [hereinafter *SOA Report*].

[92] *See, e.g.*, Rudin & Ustun, *supra* note 86, at 12 ("Specificity is the true negative rate, the fraction of negative test observations predicted to be negative.").

[93] *SOA Report*, *supra* note 91, at 44.

[94] State v. Loomis, 881 N.W.2d 749, 754 (Wis. 2016) ("A COMPAS report consists of a risk assessment designed to predict recidivism. . . . The risk assessment portion of COMPAS generates risk scores displayed in the form of a bar chart, with three bars that represent pretrial recidivism risk, general recidivism risk, and violent recidivism risk.").

[95] *Id.* ("Each bar indicates a defendant's level of risk on a scale of one to ten.").

health-care algorithm utilized by the hospital in *Dissecting Bias* also assigns risk scores with numerical values.[96] These types of ordinal values can be used to classify individuals into dichotomous groups by assigning thresholds to the underlying measure. The hospital described in *Dissecting Bias*, for example, identified very high-risk patients as those who were at the 97th percentile of risk scores.[97] The hospital utilized a second threshold, at the 55th percentile, to identify potentially at-risk patients.[98] To assess the overall predictive accuracy of algorithm, the sensitivity and specificity can be computed across all selected threshold values to generate a Receiver Operator Characteristic (ROC) curve where the area under the curve (AUC) is regarded as "an effective measure of accuracy."[99]

> The Receiver Operator Characteristic (ROC) curve is a plot of true positive rate for each possible value of the false positive rate. The area under the ROC curve (AUC) is important, since if the true positive rate is high for each value of the false positive rate, the algorithm has a high AUC and is performing well. An AUC value of .5 would be obtained for random guessing, an AUC of 1 is perfect . . . .[100]

As noted by the Society of Actuaries in its 2016 report, *Accuracy of Claims-Based Risk Scoring Models* (the SOA Report),[101] "ROC curves are typically compared by calculating the area under the curve. A perfect model would have an area under the ROC curve (AUC) of 1.0, compared to a naïve model of random guesses . . . which would have an AUC of 0.5."[102] As a general rule, an "AUC of 0.5 suggests no discrimination" (i.e., same accuracy as flipping a coin), "0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered *excellent*, and more than 0.9 is considered outstanding."[103]

As noted, empirically assessing black-box algorithms is complicated by their opacity.[104] Because the manufacturer's business practices obscure the mathematical relationship between the algorithm's inputs and outputs,[105] "its calculations cannot be double-checked for individual

[96] *See* Dan Dunn & Mark Leenay, *Value-Driven Population Health Strategies: Designing Models for Different Populations*, OPTUM (2012) ("A risk score of 1 reflects the healthiest, lowest-risk segment. The population sector with a 15 score is the most ill and has the highest propensity for services utilization.").

[97] Obermeyer et al., *supra* note 1, at 448 ("Patients above the 97th percentile are automatically identified for enrollment in the program.").

[98] *Id.* ("Those above the 55th percentile are referred to their primary care physician. . . .").

[99] Hajian-Tilaki, *supra* note 89, at 627.

[100] Rudin & Ustun, *supra* note 86, at 12.

[101] *SOA Report*, *supra* note 91.

[102] *Id.* at 44.

[103] Jayawant N. Mandrekar, *Receiver Operating Characteristic Curve in Diagnostic Test Assessment*, 5 J. THORACIC ONCOLOGY 1315, 1316 (Sept. 2010) (discussing the use of ROC curves in assessing the validity of diagnostic tests).

[104] *See* discussion *supra* Section I.B.1.

[105] *See, e.g.*, State v. Loomis, 881 N.W.2d 749, 761 (Wis. 2016) ("Northpointe, Inc., the developer of COMPAS, considers COMPAS a proprietary instrument and a trade secret. . . . [I]t does not disclose how the risk scores are determined or how the factors are weighed.").

cases, and its methodology cannot be verified."[106] Unlike transparent risk-scoring algorithms, where independent medical researchers can directly evaluate and improve the algorithm, researchers studying black-box algorithms "must work 'from the outside,' often with great ingenuity, and resort to clever work-arounds such as audit studies."[107] Jurisdictions in New York,[108] California,[109] and Florida,[110] for example, have commissioned validation studies of the COMPAS risk-scoring software, a black-box algorithm used in the criminal justice system throughout the United States. The NCSC surveyed multiple assessments and found that the AUCs of COMPAS's predictions were between 0.50 to 0.73,[111] placing the range between a coin flip and "acceptable." Acceptable is less than the excellent (0.8 to 0.9) or outstanding (more than 0.9) benchmarks used in the medical setting.[112] In the criminal justice context, however, a 0.7 is often considered acceptable or "satisfactory."[113]

What is acceptable to an institution, however, is not necessarily acceptable to the individuals adversely impacted by the algorithm's error rate. The COMPAS algorithm, for example, assigned a high recidivism risk score to an eighteen-year-old Black girl arrested for petty theft.[114] The same algorithm assigned a low risk score to a forty-one-year-old white seasoned criminal who previously served a five-year prison term for armed robbery.[115] History proved both predictions to be incorrect: the eighteen-year-old Black girl did not commit another criminal act within the predicted time period, but the forty-one-year-old white seasoned criminal did.[116] The algorithm's false positive rate adversely impacted the Black eighteen-year-old girl, and the algorithm's false negative rate adversely impacted the forty-one-year-old convict's future victims who

[106] Rudin et al., *supra* note 73, at 4.

[107] *See* Obermeyer et al., *supra* note 1, at 447.

[108] SHARON LANSING, N.Y. STATE DIV. OF CRIM. JUST. SERVS., OFF. OF JUST. RSCH. & PERFORMANCE, NEW YORK STATE COMPAS-PROBATION RISK AND NEED ASSESSMENT STUDY: EXAMINING THE RECIDIVISM SCALE'S EFFECTIVENESS AND PREDICTIVE ACCURACY i–ii (2012) (report prepared to present findings from a study examining the effectiveness and accuracy of COMPAS).

[109] JENNIFER L. SKEEM & JENNIFER E. LOUDON, U.C. DAVIS, ASSESSMENT OF EVIDENCE ON THE QUALITY OF THE CORRECTIONAL OFFENDER MANAGEMENT PROFILING FOR ALTERNATIVE SANCTIONS (COMPAS) 4–6 (2007) (report prepared for the California Department of Corrections and Rehabilitation).

[110] CTR. FOR CRIMINOLOGY & PUB. POL'Y RSCH., COLL. OF CRIMINOLOGY & CRIM. JUST., FLA. STATE UNIV., VALIDATION OF THE COMPAS RISK ASSESSMENT CLASSIFICATION INSTRUMENT 11–12 (2010) (report prepared for the Broward Sheriff's Office, Department of Community Control).

[111] Casey et al., *supra* note 84, at A-23.

[112] Mandrekar, *supra* note 103, at 1316.

[113] *See, e.g.*, LANSING, *supra* note 108, at 4.

[114] *See* Angwin et al., *supra* note 78.

[115] *Id.*

[116] *Id.*

may have been spared had the forty-one-year-old convict's score accurately reflected his actual risk.[117]

Further, even algorithms with an AUC above 0.5 may not be any more accurate than humans. A 2018 study, for example, found that humans with little or no criminal justice expertise were no less accurate at predicting recidivism than the COMPAS algorithm.[118] Under every measure, including overall accuracy and percentages of false positives and false negatives, the untrained humans' predictions were the same or better than the black-box algorithm.[119]

Empirically validating an algorithm's error rate is paramount in the medical setting. Congruent with that objective, the primary objective of the SOA Report was to "evaluate the predictive accuracy of the current set of commercial risk scoring models available in the marketplace."[120] One metric used to evaluate each algorithm was its "predictive ratio," which measures an algorithm's accuracy as a percentage equal to the predicted costs divided by the actual costs.[121] "A predictive ratio in excess of 100 percent indicates that a model overestimates the risk level for that group, while a predictive ratio below 100 percent indicates that the model underestimates the risk level."[122]

According to the SOA Report, when the algorithm was tasked with making actuarial predictions about patient cohorts based on their age and gender, the predictive ratios for the algorithm described in *Dissecting Bias* varied between 116% on the high end and 95.4% on the low end.[123] When predicting costs associated with specific conditions, however, such as heart disease, mental illness, or diabetes, the algorithms' predictive ratios fell between 59.9% on the low end and 85% on the high end.[124] When predicting costs based on cost percentile, algorithms uniformly missed the mark with predictive ratios of 8420% on the high end of the spectrum and 29% on the low end.[125] The predictive accuracy of an algorithm, therefore, is highly dependent on the type of prediction the algorithm is asked to make. Based on the foregoing reports, the algorithm

[117] *Id.* (reporting that the 41-year-old with the low risk score "subsequently br[oke] into a warehouse and st[ole] thousands of dollars' worth of electronics.").

[118] Dressel & Farid, *supra* note 74, at 1 ("We show . . . COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS's collection of 137 features, the same accuracy can be achieved with a simple linear predictor with only two features.").

[119] *Id.* at 2 tbl.1.

[120] *SOA Report*, *supra* note 91, at 5.

[121] *Id.* at 13 ("Predictive ratios are defined as the mean risk score divided by the mean actual cost for a subgroup of individuals from the sample population, with both values scaled to 1.0 over the entire population.").

[122] *Id.* at 24.

[123] *Id.* at 30 tbl.4.4.5.

[124] *Id.* at 73 tbl.I.C.1.

[125] *See id.* at 77 tbl.I.C.5.

analyzed in *Dissecting Bias* appears to be "highly predictive of cost"[126] when it is asked to predict near-term health-care costs associated with a group of patients based on their age and gender.[127] If, on the other hand, you ask about the health-care costs associated with a specific condition—such as heart disease—you may get a much less accurate answer.[128]

To assess the accuracy of predictions about individual patients, the SOA Report includes AUC assessments of the algorithms' ability to correctly identify the top 1% of the most costly patients.[129] All of the analyzed algorithms had AUCs of 0.8–0.9,[130] which is considered excellent or outstanding,[131] with the algorithm analyzed by *Dissecting Bias* having AUC values of 0.92–0.959.[132] To assess predictions about the other 99% of patients, the SOA Report examined the accuracy of the algorithms at making predictions within a "tolerable error range," which examines, for a given algorithm, "what is the probability that the model accurately predicts the risk of an individual within a certain tolerance?"[133] When, for example, the tolerable error range is set to a maximum of 100% of the actual scaled costs, all of the prospective model algorithms have AUCs of approximately 50%, with the algorithm analyzed by *Dissecting Bias* having an AUC of 52.2%.[134] When the tolerable error range is increased to 300% of costs, the accuracy of the algorithm's predictions increases significantly, with an AUC range of 75–80%.[135]

In *Dissecting Bias*, however, the hospital was not using the algorithm to predict costs.[136] Rather, it was using the algorithm's advertised ability to "flag individuals" who have multiple chronic conditions and determine "before their health becomes catastrophic" and "in need of specialized intervention programs."[137] Even if the algorithm is "highly predictive of cost" because that "is what it was designed to do,"[138] the analyses reflected in the SOA Report do not quantify the algorithms' abilities to predict specific medical conditions underlying those costs.[139] Moreover, as indicated in the results described above, the algorithm's

[126] Snowbeck, *supra* note 15.

[127] *See SOA Report*, *supra* note 91, at 30 tbl.4.4.5.

[128] *See id.* at 77 tbl.I.C.5.

[129] *Id.* at 43–46.

[130] *Id.* at 45 fig.4.

[131] Mandrekar, *supra* note 103, at 1316.

[132] *SOA Report*, *supra* note 91, at 45 fig.4.

[133] *See id.* at 40.

[134] *Id.* at 4 tbl.4.5.3.

[135] *Id.*

[136] *See Supplementary Materials*, *supra* note 20, at 3 ("The algorithm's stated goal (from promotional materials) is to predict which individuals are in need of specialized intervention programs and which intervention programs have the most impact on the quality of individuals' health.").

[137] *Id.*; *see also* OPTUM, *supra* note 20 (containing the advertised abilities of the algorithm).

[138] Snowbeck, *supra* note 15 (quoting UnitedHealth Group statement).

[139] *See SOA Report*, *supra* note 91, at 16–48 (describing the results of the study of multiple claims-based risk scoring models).

reliability at predicting health-care costs varies depending on the specific questions being asked.[140] On average, health-care costs may be strongly correlated with medical needs, as noted by *Dissecting Bias*.[141] Merely being aware that a correlation exists, however, is not a measure of that correlation, nor is that general awareness equivalent to a numerical value required to calculate how many false positives and false negatives the algorithm is likely to produce. As noted above,[142] the ability of the algorithm to predict the costs of specific individuals varies significantly depending on the tolerance range, with an AUC of over 95% in some situations,[143] and an AUC of as low as 52.2% in others.[144]

3. Assessing Bias

There are many mechanisms by which biases can inadvertently influence the development or operation of an algorithm which can result in "disproportionate[] adverse outcomes" for minority groups.[145] "This is because," according to Professors Prince and Schwarcz, "AIs will affirmatively 'seek out' proxies for prohibited, but predictive, characteristics within increasingly vast amounts of training data."[146] Many state laws, for example, prohibit insurers from discriminating against victims of domestic violence notwithstanding the fact that such victims will frequently present greater financial risks to the insurers.[147] Some life insurance companies have, therefore, looked to AI to achieve the same outcomes by "predicting life expectancy by relying on proxies that derive from social media."[148] Rather than directly looking at whether the individual is a victim of domestic violence, the AI can look for proxies such as the "websites an individual visits, the location and information in their cell phones, or their social media posts."[149] Thanks to the training data, the development process "produce[s] the very same results that the law sought to avoid."[150]

The algorithm analyzed in *Dissecting Bias* specifically excludes the category "race" as an input to its calculations.[151] The algorithm uses the information that is otherwise submitted to an insurer in connection with

[140] *See supra* text accompanying notes 100–12.

[141] Obermeyer et al., *supra* note 1, at 450 ("Health care costs and health needs are highly correlated, as sicker patients need and receive more care, on average.").

[142] *See supra* text accompanying notes 99–12.

[143] *SOA Report*, *supra* note 91, at 45 fig.4.

[144] *Id.* at 43, tbl.4.5.3.

[145] *Big Data*, *supra* note 3, at 671, 677.

[146] Prince & Schwarcz, *supra* note 3, at 1275–76.

[147] *Id.* at 1294.

[148] *Id.* at 1295.

[149] *Id.* at 1294.

[150] *Id.*

[151] Obermeyer et al., *supra* note 1, at 449 ("Notably, the algorithm specifically excludes race."); *see also* Obermeyer et al., *supra* note 35, at 3 ("For the commercially insured sample . . . specifically exclud[ed] race.").

processing a patient's medical claim.[152] Insurers do not receive medical claims from patients who do not obtain medical services.[153] Consequently, "there are many opportunities for a wedge to creep in between needing health care and receiving health care. . . ."[154] One potential wedge noted by the authors is poverty, which "can lead to disparities in use of health care: geography and differential access to transportation, competing demands from jobs or child care, or knowledge of reasons to seek care."[155] "To the extent," therefore, "that race and socioeconomic status are correlated, these factors will differentially affect Black patients."[156]

Inequality in Quality, Addressing Socioeconomic, Racial, and Ethnic Disparities in Health Care[157] summarizes multiple studies indicating health-care disparities experienced by non-white people:

> Elderly [B]lacks, compared with whites, are seen less often by specialists, receive less appropriate preventive care including mammography and influenza vaccinations, lower quality hospital care, and fewer expensive, technological procedures. In general, [B]lacks receive less intensive hospital care, including fewer cardiovascular procedures, lung resections for cancer, kidney and bone marrow transplants, cesarean sections, peripheral vascular procedures, and orthopedic procedures. They have also been reported to receive less aggressive treatment of prostate cancer, fewer antiretrovirals for human immunodeficiency virus infection, antidepressants for depression, tympanostomy tubes, and admissions for chest pain, and lower-quality prenatal care.[158]

Dissecting Bias found that the claims data utilized by the algorithm in the study reflected similar disparities:[159]

> At a given level of health (again measured by number of chronic illnesses), Blacks generate lower costs than [w]hites—on average, $1801 less per year, holding constant the number of chronic illnesses (or $1144 less, if we instead hold constant the specific individual illnesses that contribute to the sum).[160]

[152] Obermeyer et al., *supra* note 1, at 449 ("In our setting, the algorithm takes in a large set of raw insurance claims data . . . demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs."); *see also SOA Report*, *supra* note 91, at 10 ("Uses information readily available from medical and pharmacy claims . . . . Uses a member's clinical episodes of care, prior use of health care services, prescription drugs, and lab results as markers of their future health care use.").

[153] Disregarding the comparatively rarer instances of medical fraud.

[154] Obermeyer et al., *supra* note 1, at 450.

[155] *Id.*

[156] *Id.*

[157] Kevin Fiscella, Peter Franks, Marthe R. Gold, & Carolyn M. Clancy, *Inequality in Quality, Addressing Socioeconomic, Racial, and Ethnic Disparities in Health Care*, 283 J. AMER. MED. ASS'N No. 19, 2579 (May 17, 2000), https://jamanetwork.com/journals/jama/article-abstract/192714.

[158] *Id.* at 2579–80 (citing approximately two dozen studies).

[159] *See* Obermeyer et al., *supra* note 1, at 450.

[160] *Id.*

These disparities appeared to be reflected in the algorithm's predictions, in which Black patients receiving the same risk score had "26.3% more chronic illnesses than [w]hites."[161] This, in turn, could impact patient care, like when the hospital in *Dissecting Bias* identified patients for enrollment into the enhanced-care coordination program if their risk scores were above the 97th percentile.[162] Using the algorithm's risk scores as a measure of health-condition severity, only 17.7% of patients in the top 3rd percentile were Black; if, however, the hospital used the study's independently developed measure of comorbidity, that percentage would rise to 46.5%.[163] As a method for identifying the top 3rd percentile of the most at-risk patients, the algorithm had a significant number of false positives that were disproportionately white patients, and false negatives that were disproportionately Black patients.[164] Additionally, the developer of the algorithm replicated this discrepancy on a national dataset of 3.7 million patients, finding that "Black patients had 48,772 more active chronic conditions than [w]hite patients, conditional on risk score. . . ."[165]

Based on these results, the *Dissecting Bias*'s authors found that this "widely used algorithm . . . exhibits significant racial bias."[166] This raises the question of how do these findings square with the developer's statement that the algorithm is "highly predictive of cost, which is what it was designed to do?"[167] As previously discussed, the algorithm's accuracy varies with the context in which it is made.[168] Its predictions performed very well at correctly identifying the top 1% of most costly patients, with an AUC of 0.995,[169] and performed less well when predicting a patient's medical costs with a 100% tolerable error range and an AUC of 0.522.[170] Notably, however, *Dissecting Bias* found no evidence of bias when it was used to predict cost, which is "what it was designed to do."[171] "By this metric," the authors note, "one could call the algorithm unbiased. . . . [A]t every level of algorithm-predicted risk, Blacks and [w]hites have

[161] *Id.* at 448.

[162] *Id.* ("Patients above the 97th percentile are automatically identified for enrollment in the program.").

[163] *Id.* at 449 ("[A]t α = 97th percentile, among those auto-identified for the program, the fraction of Black patients would rise from 17.7 to 46.5%.").

[164] *See id.* at 448–49 fig.1 (demonstrating the fraction of patients at or above a given risk score).

[165] *Id.* at 453 ("[T]he manufacturer independently replicated our analyses on its national dataset of 3,695,943 commercially insured patients . . . [finding that] . . . Black patients had 48,772 more active chronic conditions than White patients, conditional on risk score. . . .").

[166] *Id.* at 448.

[167] Snowbeck, *supra* note 15.

[168] *See* discussion *supra* Section I.B.2.

[169] *SOA Report*, *supra* note 91, at 45 fig.4.

[170] *Id.* at 43 tbl.4.5.3.

[171] Snowbeck, *supra* note 15.

(roughly) the same costs the following year. . . . Conditional on risk score, predictions do not favor [w]hites or Blacks. . . ."[172]

These results indicate that the bias observed in *Dissecting Bias* may not have been inherent in the development of the algorithm, but rather could have arisen from the use of the algorithm's risk scores as a proxy for identifying patients' medical needs. If so, this appears to be a use touted in the developer's marketing materials that talk about the algorithm's ability to "determine which individuals are in need of specialized intervention programs," "[f]lag individuals for intervention before their health becomes catastrophic," and "identify members with upcoming evidence-based medicine gaps in care for early engagement."[173]

C. Mitigating Accuracy and Bias Risks

1. Mitigating Algorithmic Inaccuracy

HIPAA's security rule requires health-care organizations to "[e]nsure the confidentiality, integrity, and availability of all [patient information that it] creates, receives, maintains, or transmits."[174] Health-care organizations must "[c]onduct an accurate and thorough assessment of the potential risks and vulnerabilities to the . . . integrity" of the health information they hold.[175] Under HIPAA's security rule, "integrity" usually pertains to assurances that patient information is not "altered or destroyed in an unauthorized manner."[176] With respect to the use of predictive models, the concept of "integrity" could be adapted to protect against the reasonably anticipated threats[177] to the reliability risk scores described above.[178]

The output of *Dissecting Bias*'s algorithm were risk scores for each of the hospital's patients.[179] As discussed, the predictive accuracy of the risk scores varied significantly depending on the context in which they are used,[180] even when predicting costs, which, according to the developer, is what they were designed to do.[181] An "accurate and thorough as-

[172] Obermeyer et al., *supra* note 1, at 450–51.

[173] *See, e.g.*, OPTUM, *supra* note 20.

[174] 45 C.F.R. § 164.306(a)(1) ("Covered entities and business associates must . . . [e]nsure the . . . integrity . . . of all electronic protected health information the covered entity or business associate creates, receives, maintains, or transmits.").

[175] 45 C.F.R. § 164.308(a)(1)(ii)(A).

[176] 45 C.F.R. § 164.304 ("*Integrity* means the property that data or information have not been altered or destroyed in an unauthorized manner.").

[177] 45 C.F.R. § 164.306(a)(2) ("Covered entities and business associates must . . . [p]rotect against any reasonably anticipated threats or hazards to the . . . integrity of such information.").

[178] *See* discussion *supra* Section I.B.

[179] Obermeyer et al., *supra* note 1, at 449 ("In the health system we studied, risk scores are generated for each patient. . . .").

[180] *See supra* text accompanying notes 100–12.

[181] Snowbeck, *supra* note 15 ("[T]he cost model within Impact Pro was highly predictive of cost, which is what it was designed to do.") (quoting Optum).

sessment of the potential risks"[182] to the reliability of a risk score, therefore, would recognize that the risk score's predictive accuracy is highly sensitive to the context in which it is used.[183]

When a risk score is marketed for purposes other than "what it was designed to do,"[184] such as to "identify [individuals] with upcoming evidence-based medicine gaps in care for [proactive engagement]"[185] the risk to the predictive reliability of the scores is undiminished. Healthcare costs may be strongly correlated with medical needs because on average, "sicker patients need and receive more care. . . ."[186] At the same time, "there are many opportunities for a wedge to creep in between needing health care and receiving health care."[187] Merely being aware that a correlation exists is not a measure of that correlation that can be used to calculate the number of false positives or false negatives the algorithm is likely to generate.

Dissecting Bias provides limited information about the hospital's implementation of the risk scores:

> In the health system we studied, risk scores are generated for each patient during the enrollment period for the system's care management program. Patients above the 97th percentile are automatically identified for enrollment in the program. Those above the 55th percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment.[188]

The passage does not describe the hospital's motivation for this two-tiered approach where patients above the 97th percentile are automatically identified for enrollment in the program, and those above the 55th percentile are referred to a primary care physician.[189] These thresholds do, however, roughly correlate to the algorithm's predictive reliability described in the SOA Report.[190] As previously discussed, those risk scores performed well at correctly identifying the top 1% of most costly patients, with an AUC of greater than 0.92.[191] This roughly corresponds to the hospital's use of the 97th percentile as the threshold for patients

[182] 45 C.F.R. § 164.308(a)(1)(ii)(A).

[183] *See supra* text accompanying notes 100–12.

[184] Snowbeck, *supra* note 15 ("[T]he cost model within Impact Pro was highly predictive of cost, which is what it was designed to do.") (quoting Optum).

[185] OPTUM, *supra* note 20; *see also* Obermeyer et al., *supra* note 35, at 3 ("The algorithm's stated goal (from promotional materials) is to predict which individuals are in need of specialized intervention programs and which intervention programs have the most impact on the quality of individuals' health.") (citing the developer's promotional materials).

[186] Obermeyer et al., *supra* note 1, at 451.

[187] *Id.*

[188] *Id.* at 449.

[189] *Id.*

[190] *See supra* text accompanying notes 100–12.

[191] *SOA Report*, *supra* note 91, at 45 fig.4.

who are automatically identified for enrollment in the care management program.[192] The algorithm's risks scores, however, were far less predictive for other groups of patients, with an AUC of 0.522, when predicting a patient's medical costs with a 100% tolerable error range.[193] This degradation of predictive accuracy may have motivated the hospital's use of primary care physicians to evaluate whether patients above the 55th percentile would benefit from the care management program, as well as furnishing the doctor's contextual information with the patient's electronic health records and insurance claims.[194] Additionally, the degradation of predictive accuracy may have motivated the hospital's selection of the 55th percentile. Studies have indicated that approximately 25%–38% of American adults have multiple chronic conditions.[195] The 55th percentile threshold, therefore, operates as a buffer against the diminished predictive accuracy of the algorithm in identifying patients outside the top 1% of the costliest patients.

2. Mitigating Algorithmic Bias

As previously discussed, there is significant risk that an algorithm highly predictive of cost[196] can exhibit significant racial bias whenever hospitals use its risk scores to predict patients' medical needs.[197] The manufacturer of the algorithm appeared to acknowledge as much in its response to *Dissecting Bias*'s publication, stating "[t]hese gaps, often caused by social determinants of care and other socio-economic factors, can then be addressed by the health systems and doctors to ensure people, especially in underserved populations, get effective, individualized care."[198]

Acknowledgement of a risk, however, is distinct from an effective plan to mitigate the potential harm. HIPAA's security rule, for example, clearly distinguishes the process of assessing potential risks[199] from the process of reducing identified risks to an appropriate level.[200] As discussed above, the hospital in *Dissecting Bias* referred patients with risk scores in the top 55th percentile to a physician, potentially to mitigate

[192] Obermeyer, *supra* note 1, at 449.

[193] *SOA Report*, *supra* note 91, at 43 tbl.4.5.3.

[194] Obermeyer et al., *supra* note 1, at 453 ("Specifically, for patients at or above [the 55th percentile] doctors are presented with contextual information from patients' electronic health records and insurance claims and are prompted to consider enrolling them in the program.").

[195] *See, e.g.*, Brian W. Ward & Lindsey I. Black, *State and Regional Prevalence of Diagnosed Multiple Chronic Conditions Among Adults Aged ≥18 Years — United States, 2014*, 65 MORBIDITY AND MORTALITY WKLY. REP. 735, 735 (Jul. 29, 2016), https://www.cdc.gov/mmwr/volumes/65/wr/mm6529a3.htm.

[196] Snowbeck, *supra* note 15.

[197] *See* discussion *supra* Section I.B.3.

[198] Snowbeck, *supra* note 15.

[199] 45 C.F.R. § 164.308(a)(1)(ii)(A) ("Risk analysis . . . [c]onduct[ing] an accurate and thorough assessment of the potential risks and vulnerabilities. . . .").

[200] 45 C.F.R. § 164.308(a)(1)(ii)(B) ("Risk management . . . [i]mplement[ing] . . . measures sufficient to reduce risks and vulnerabilities to a reasonable and appropriate level. . . .").

concerns over the predictive accuracy of risk scores outside the top 1% of predicted costs.[201] This introduces the question of whether the intervention of physicians in the process mitigates the risk of bias in determining which patients are enrolled in the care management program.

Dissecting Bias offered a limited answer to this question by comparing the racial composition of sample patients that the hospital enrolled in the care management program to simulations of what the implemented algorithm would predict.[202] In this experiment, the percentage of Black patients who would be enrolled in the program based solely on their risk scores would be 17.2%.[203] This is lower than the percentage of Black patients actually enrolled by physicians, which was 19.2%,[204] suggesting that physician engagement may offset some of the racial bias reflected in the risk scores generated by the algorithm. At the same time, *Dissecting Bias* noted that, had patients been enrolled based solely on the number of their active chronic conditions, 29.2% of those patients would have been Black,[205] a significantly higher percentage than those actually enrolled. From this, the authors conclude that "although doctors do redress a small part of the algorithm's bias, they do so far less than" using an algorithm based on patients' medical conditions.[206]

CONCLUSION: FOSTERING ACCOUNTABILITY

A. Organizational Accountability

Organizations that elect to automate their decision-making through AI algorithms have options. They can, for example, choose to implement black-box algorithms without understanding how they work, without obtaining independent validation of the algorithms' sensitivity and specificity, or without recognizing the likelihood of significant racial bias. This cavalier approach, however, has consequences. It can readily serve to perpetuate the use of unreliable black-box algorithms by allowing inaccuracies and biases to persist indefinitely due in large part to the tolerance for opacity. This, in turn, can stunt the tremendous promise of AI to serve as an agent of improved accuracy and fairness.

Organizations could choose to take a more proactive approach by accepting responsibility for the accuracy of the decisions they delegate to algorithms. For health-care organizations, this would require a relatively minor adaptation of their current health information integrity risk-

[201] *See* discussion *supra* Section I.C.1.

[202] Obermeyer et al., *supra* note 1, at 453 (discussion under "Relation to human judgment").

[203] *Id.* ("[W]e compare this to simply assigning those with the highest predicted costs . . . which would yield 17.2.").

[204] *Id.* ("The enrolled individuals are 19.2% Black.").

[205] *Id.* ("[W]e compare this to simply assigning those with . . . the highest number of active chronic conditions, to the program . . . which would yield . . . 29.2% Black patients.").

[206] *Id.* ("Thus, although doctors do redress a small part of the algorithm's bias, they do so far less than an algorithm trained on a different label.").

management process. Before integrating the outputs of any decision-making algorithm, they would conduct an accurate and thorough risk assessment of the algorithm. This would include a rigorous understanding of the algorithm's predictive accuracy with respect to the specific intended uses of the outputs, including having reasonably accurate assessments of their sensitivity and specificity. To the extent that the algorithm concentrates its error rate in a manner that adversely impacts one or more communities, the health-care organization can estimate that impact. These assessments, in turn, would inform the organization's implementation through the adoption of safeguards designed to mitigate the impact of the risks identified in the risk assessment.

Second, health-care organizations can foster the adoption of these practices by publicly sharing their risk assessment and mitigation methodologies. Such transparency can help identify organizational blind spots and validate approaches, such as assessing the circumstances in which human intervention is effective at counteracting observed biases in an algorithm's risk scores. Publicly validated approaches can, in turn, serve as the bedrock for industry standards and regulatory requirements.

B. Developer Accountability

If customer risk-management processes become ubiquitous when implementing decision-making algorithms, this could create an environment where developers start competing on transparency and accuracy. Health-care organizations selecting from a menu of available algorithms may find it less risky to use transparent algorithms whose accuracy can be directly validated by customers, rather than proprietary black-box algorithms where "researchers must work 'from the outside,' often with great ingenuity, and resort to clever workarounds" to objectively assess the output's accuracy.[207]

One rationale for maintaining the black-box nature of algorithms is found in *Loomis v. State*,[208] where the "proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined."[209] Developers could market shrouding proprietary algorithms behind a veil of complexity, such as requiring 137 variables to make a prediction, as a value proposition to customers.[210] Journalists have reported, for example, that Broward County pays over $20,000 per year to use its black-box criminal justice algorithm.[211] Would Broward County pay $20,000 per year to use

[207] *Id.* at 448.

[208] 881 N.W.2d 749 (Wis. 2016).

[209] *Id.* at 769.

[210] *See* Rate of Change Podcast, *Opening the Black Box*, DUKE UNIV. PRATT SCH. ENG'G (Aug. 30, 2019), https://pratt.duke.edu/about/news/podcast/opening-black-box (discussing how the 137 factors COMPAS uses are as accurate as professor's own code).

[211] Angwin et al., *supra* note 78.

the simple rule based on two or three variables, such as age, gender, and criminal history? Preserving the proprietary interests of developers, however, is not a benefit for either the individuals adversely impacted by an algorithm's errors or customers who need as much transparency as possible to mitigate the harm such errors may cause.

A second justification derives from the "widely held belief that there is a tension between how well a model will perform and how well humans will be able to interpret it."[212] The ability of AI to create models that are indecipherable to humans has arisen alongside a widespread presumption that indecipherable models are often more accurate than those humans are able to comprehend:[213]

> The power of machine learning lies not only in its ability to relieve programmers of the difficult task of producing explicit instructions for computers, but in its capacity to learn subtle relationships in data that humans might overlook or cannot recognize. This power can render the models developed with machine learning exceedingly complex and, therefore, impossible for a human to parse.[214]

"This view reflects the reasonable idea that models that consider a larger number of variables, a larger number of relationships between these variables, and a more diverse set of potential relationships is likely to be *both* more accurate and more complex."[215]

According to Professors Cynthia Rudin and Joanna Radin, however, there has been little evidence supporting this presumption: "In the criminal justice system, it has been repeatedly demonstrated . . . that complicated black box models for predicting future arrest are not any more accurate than very simple predictive models based on age and criminal history."[216]

An example of this is provided in *Certifiably Optimal Rule Lists for Categorical Data*,[217] where researchers used interpretable machine learning to develop a criminal recidivism model that was equally as predictive as COMPAS, the black-box algorithm that is widely used throughout the United States.[218] Moreover, rather than requiring 137 inputs to produce a prediction, the accountable AI model utilizes three inputs readily under-

[212] *Explainable Machines*, *supra* note 3, at 1110.

[213] 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*, HARV. DATA SCI. REV. 1.2, 2 (Nov. 22, 2019) ("Over the last few years, the advances in deep learning for computer vision have led to a widespread belief that the most accurate models for any given data science problem must be inherently uninterpretable and complicated.").

[214] *Explainable Machines*, *supra* note 3, at 1094.

[215] *Id.* at 1110.

[216] Rudin & Radin, *supra* note 217, at 4.

[217] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, & Cynthia Rudin, *Learning Certifiably Optimal Rule Lists for Categorical Data*, J. MACH. LEARNING RSCH. 18 (2018).

[218] *Id.*

standable by any user: an individual's age, gender, and criminal history.[219]

> The full machine learning model is as follows: if the person has either >3 prior crimes, or is 18–20 years old and male, or is 21–23 years old and has two or three prior crimes, they are predicted to be rearrested within two years from their evaluation, and otherwise not . . . this set of rules is as accurate as the widely used (and proprietary) black box model called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). . . .[220]

The simplicity of this model has enormous benefits to the organizations that use it. If, for example, software running this model assigns a high recidivism risk score to an eighteen-year-old Black girl arrested for petty theft,[221] any judge reviewing this risk score can compare it to the inputs and be alerted to the erroneous result. Judges would also understand that there is no "magic" to the risk score, that even this model has an error rate, and that they should not uncritically turn a blind eye to facts that suggest the algorithm's prediction may be inaccurate.

The same methodology applies to health-care. Professors Rudin and Radin cite studies where researchers were able to find simple, interpretable algorithms to predict pneumonia risk and[222] Type 2 diabetes,[223] noting:

> There also does not seem to be a benefit in accuracy for black box models in several healthcare domains and across many other high-stakes machine learning applications where life-altering decisions are being made (. . . who all show models with interpretability constraints that perform just as well as unconstrained models).[224]

In performing their analysis, the authors of *Dissecting Bias* took steps to develop a transparent alternative to determining eligibility to participate in the hospital's enhanced care coordination services.[225] Rather than purporting to predict whether a patient had multiple chronic conditions, researchers used the patient's electronic medical records to create a "comorbidity score" that sought to measure the number of active

[219] *See* Rudin & Radin, *supra* note 217, at 4.

[220] *Id.*

[221] *See* Angwin et al., *supra* note 78.

[222] Rudin & Radin, *supra* note 213, at 8 (citing Caruana et al., *supra* note 10).

[223] *Id.* (citing Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, & David Sontag, *Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors*, 3 BIG DATA 4, 277–87 (2015)).

[224] *Id.* at 4.

[225] Obermeyer et al., *supra* note 1, at 449 ("We begin by calculating an overall measure of health status, the number of active chronic conditions [or 'comorbidity score,' a metric used extensively in medical research . . . to provide a comprehensive view of a patient's health]"); *see also Supplementary Materials*, *supra* note 20, at 4–5 (discussing the development of the study's comorbidity score).

chronic conditions a patient had,[226] defined as "how many chronic conditions are flaring up . . . not simply an indicator of previously diagnosed chronic conditions."[227] By focusing on *active* chronic conditions, rather than *previously diagnosed* conditions, the measure sought to prioritize patients based on whether the conditions presented a current risk to the patients.[228]

The researchers in *Dissecting Bias* used the chronic comorbidity score as a tool solely to assess the cost-based algorithm.[229] However, it demonstrates how efficiently alternatives to black-box algorithms can be developed in algorithmic decision-making.[230] As noted by Professors Rudin and Radin, "[t]he belief that accuracy must be sacrificed for interpretability is inaccurate."[231] Health-care organizations looking to implement accurate and unbiased algorithms to support their decision-making, therefore, can elect to work with developers whose algorithms are transparent, rather than black boxes. This could, in turn, significantly streamline the risk assessments and mitigation planning they would otherwise conduct to ensure that the decisions they delegate to these algorithms are as accurate as they presume and do not adversely impact the delivery of care.

[226] Obermeyer et al., *supra* note 1, at 449.

[227] *Supplementary Materials*, *supra* note 20, at 6. "Our goal was to construct biomarker-based measures of severity for as many of these illnesses as possible. This was meant to measure not just the presence or absence of these illnesses, but the degree to which they are well managed. . . ." *Id.* at 4. "Of note, this is a measure of how many chronic conditions are flaring up and driving utilization, not simply an indicator of previously diagnosed chronic conditions (for which predictions are not necessarily required)." *Id.* at 6.

[228] *Id.*

[229] *Id.*

[230] *Id.*

[231] Rudin & Radin, *supra* note 213, at 3.