Articles

Algorithmic Discrimination Is an Information Problem

IGNACIO N. COFONE[†]

While algorithmic decision-making has proven to be a challenge for traditional antidiscrimination law, there is an opportunity to regulate algorithms through the information that they are fed. But blocking information about protected categories will rarely protect these groups effectively because other information will act as proxies. To avoid disparate treatment, the protected category attributes cannot be considered; but to avoid disparate impact, they must be considered. This leads to a paradox in regulating information to prevent algorithmic discrimination. This Article addresses this problem. It suggests that, instead of ineffectively blocking or passively allowing attributes in training data, we should modify them. We should use existing pre-processing techniques to alter the data that is fed to algorithms to prevent disparate impact outcomes. This presents a number of doctrinal and policy benefits and can be implemented also where other legal approaches cannot.

[†] Assistant Professor, McGill University, Faculty of Law. ignacio.cofone@mcgill.ca. I'm indebted to Ian Ayres, Jack Balkin, Miriam Buiten, Richard Gold, Claudia Haupt, Ido Kilovaty, Joshua Kroll, David Lehr, Amanda Levendowski, Yafit Lev-Aretz, Asaf Lubin, Mark MacCarthy, Ethan Macdonald, John Nay, Helen Nissenbaum, Madelyn Sanfilippo, Andrew Selbst, Julia Powles, Adriana Robertson, Clare Ryan, Colleen Sheppard, Katherine Strandburg, Felix Wu, and Angela Zorro Medina for their helpful comments to different versions of this piece. This Article also benefited from comments received at the Privacy Law Scholars Conference and internal presentations at the NYU Privacy Research Group and the Yale Law School Information Society Project. I also thank Fonds de Recherche du Québec for their financial support, the editors at the Hastings Law Journal for their excellent editorial work, and Ana Qarri and Michael Beauvais for their outstanding research assistance.

TABLE OF CONTENTS

INTRODUCTION

Algorithmic decisions affect everyone, often without their knowledge. Increasingly, algorithms make impactful decisions for people's daily lives, from determining whether someone will receive a loan,[1] to determining the type of healthcare a person will receive,[2] to predicting whether someone should be granted parole.[3] While, thirty years ago, decisions that shaped people's lives were made by other people, key decisions in finance, criminal law, employment, health, politics, and online speech, to name just a few, are increasingly made by machines.[4] For example, for the question "will this candidate be a good borrower?" an algorithm compares his or her characteristics with the characteristics of those who have paid their debts. For "will this person recidivate if given parole?" it compares his or her characteristics with those that recidivated during parole.

Algorithmic decision-making offers multiple benefits to society.[5] For many tasks, algorithms surpass human abilities, and the set of those tasks is constantly expanding.[6] Given their potential, the ideal is thus not to ban algorithms, but to regulate them appropriately.[7] To help us learn how to do so, this Article focuses on one of the central promises that comes with a change from human to algorithmic decision-makers: a decrease in the prevalence of bias and discrimination.[8] The promise that is often attached to these systems is that,

---

1. FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 4–6 (2015).

2. HANNAH FRY, HELLO WORLD: BEING HUMAN IN THE AGE OF ALGORITHMS 109–12 (2018).

3. *Id.* at 54–56.

4. AJAY AGRAWAL ET AL., PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE 1–5 (2018); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 9–10 (2018).

5. Benjamin Alarie, *The Path of the Law: Towards Legal Singularity*, 66 U. TORONTO L.J. 443, 450–51 (2016).

6. Vasant Dhar, *Should We Regulate Digital Platforms?*, 5 BIG DATA 277 (2017); Vasant Dhar, *When to Trust Robots with Decisions, and When Not to*, HARV. BUS. REV. (May 17, 2016), https://hbr.org/2016/05/when-to-trust-robots-with-decisions-and-when-not-to. There seems to be a limitation to this principle: those processes that are automated within ourselves. If you ask a computer and a human to make a simple algebra calculation such as the square root of eighty-seven, the computer will obviously be able to do this faster. But if you grab and throw a ball and ask a human and a robot designed for this purpose to catch it, you will find that no robot exists so far that can do this as well as an average person. The distinction between one process and the other is normally referred to as a distinction between System 1 and System 2. *See* Keith E. Stanovich & Richard F. West, *Individual Differences in Reasoning: Implications for the Rationality Debate*, 23 BEHAV. BRAIN SCI. 645, 658 (2000); Jonathan St. B. T. Evans, *Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition*, 59 ANN. REV. PSYCHOL. 255 (2008); *see also* IN TWO MINDS: DUAL PROCESSES AND BEYOND, (Jonathan St. B. T. Evans & Keith Frankish eds., 2009).

7. *See* Andrew D. Selbst, *A Mild Defense of Our New Machine Overlords*, 70 VAND. L. REV. EN BANC 87, 87–89 (2017) (arguing that we need a realistic picture of what humans and machines can accomplish, which includes being aware of machines' defects but also seeing where machines can improve human decision-making). *Cf.* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 92–104 (2017) (describing the advantages, disadvantages, and likely future implications using algorithms).

8. CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 1–13 (2016).

while bias and prejudices plague us humans, algorithms are impartial, fair, and unbiased decision-makers.[9]

This promise has gone unfulfilled. Researchers continue to find that algorithms disproportionately disadvantage members of vulnerable minorities.[10] This makes it crucial to determine when algorithmic decisions are discriminatory and, especially given concerns about the effectiveness of traditional remedies that flow from discrimination, explore how to prevent their discriminatory outcomes. Given the increasing prevalence of algorithms in socially significant decisions, making progress on this front stands to change the daily lives of thousands of citizens.

Traditional antidiscrimination law deals with the regulation of (human) behavior around the *use* of information about others as the basis for discriminatory practices.[11] There is, however, benefit from additionally regulating the *acquisition* of such information.[12] Information rules, or privacy rules, that prevent a decision-maker from knowing the potentially discriminatory information, can prevent them from discriminating in the first place.[13]

This Article explores the relevance of this idea for artificial intelligence (A.I.) discrimination.[14] There is a growing body of literature that examines the social problems generated by algorithmic bias, the ethical dilemmas introduced by A.I., and whether the use of A.I. should be limited against disadvantaged groups,[15] but the role that the law can take in shaping how these systems are formed to reduce such harm remains under-explored. While a rich body of legal literature to date has focused on the potential harms that the increased reliance on A.I. poses, little has been said about how the law can *prevent* these harms.[16]

---

9. *But see infra* Subpart I.B.

10. *See, e.g.*, EUBANKS, *supra* note 4 at 9; SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 1–3 (2018); *see also* Mary Madden et al., *Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans*, 95 WASH. U. L. REV. 53, 64–67 (2017) (making a similar argument for big data).

11. Ignacio N. Cofone, *Antidiscriminatory Privacy*, 72 SMU L. REV. 139, 140–141 (2019).

12. Jessica L. Roberts, *Preempting Discrimination: Lessons from the Genetic Information Nondiscrimination Act*, 63 VAND. L. REV. 439, 483–84 (2010) [hereinafter *Preempting Discrimination*]; Jessica L. Roberts, *Protecting Privacy to Prevent Discrimination*, 56 WM. & MARY L. REV. 2097, 2147–56 (2014) [hereinafter *Protecting Privacy*]. *See* Anupam Datta et al., Correspondences Between Privacy and Nondiscrimination: Why They Should be Studied Together (2018), https://arxiv.org/pdf/1808.01735.pdf (showing that key aspects of privacy and nondiscrimination mirror each other at a formal level).

13. *See generally* Cofone, *supra* note 11.

14. Here, I refer to machine learning discrimination and A.I. discrimination indistinctly. *See* Ignacio N. Cofone, *Servers and Waiters: What Matters in the Law of A.I.*, 21 STAN. TECH. L. REV. 167, 179 (2018) (arguing that, for the problem of non-contractual harms, the law should treat robots, A.I. agents, and algorithms indistinctly).

15. *See, e.g.*, Melissa Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, 56 AM. CRIM. L. REV. (forthcoming 2019); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017); Madden et al., *supra* note 10.

16. Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 587 (2018). Amanda Levendowski has addressed this effort in the cited article, exploring how copyright law can be used for these purposes. *See id.* at 619–30 (arguing that copyright law is the most powerful branch of law impacting AI bias and offering solutions through the doctrine of fair use); *see also*

In particular, as David Lehr and Paul Ohm write, "almost all of the significant legal scholarship to date has focused on the implications of the running model . . . and has neglected most of the possibilities and pitfalls of playing with the data."[17] This proliferation of scholarship has led the A.I. Now Institute to state that "[t]he question is no longer whether there are harms and biases in A.I. systems. That debate has been settled: the evidence has mounted beyond doubt . . . . [t]he next task now is addressing these harms."[18] In a similar vein, a White House report from 2016 called for "manag[ing A.I.'s] risks and challenges . . . [in] ensur[ing] that everyone has the opportunity . . . to participate in its benefits,"[19] and called for "equal opportunity by design."[20]

Recent literature has shown that algorithmic discrimination works differently from purely human discrimination.[21] However, while it is true that human and algorithmic discrimination differ in many ways, they have a similar underlying information dynamic. The crucial task in preventing discrimination through information rules in both cases is identifying not only information about the protected category, but also information that acts as proxies for the protected category.[22] Antidiscriminatory information rules must focus on the information that can be used as proxies to shift discrimination to other groups,[23] both for human and machine discrimination.

For both forms of discrimination, information rules can offer short-term protection from discrimination by altering the data that the law deems harmful to use in a decision-making process. In this way, information rules can aid antidiscriminatory efforts. However, this altering of the data works differently for algorithms than it does for humans.[24] For algorithms, the solution is neither more data nor less data. It is more meaningful data. And more meaningful data means, counterintuitively, a data sample that is unrepresentative of the pool, because it looks like what we believe the pool *would* look like had it not embedded structural inequalities.

Because of the difficulties that it poses to antidiscrimination law, antidiscriminatory information rules are particularly useful to address

---

Matthew T. Bodie et al., *The Law and Policy of People Analytics*, 88 U. COLO. L. REV. 961, 1007–14 (2017). *See generally* Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM L. REV. 633 (2017).

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

18. MEREDITH WHITTAKER ET AL., AI NOW REPORT 2018 42 (2018).

19. NAT'L SCI. & TECH. COUNCIL, EXEC. OFFICE OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (Oct. 2016).

20. EXEC. OFFICE OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS (May 2016).

21. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 673–76 (2016); Pauline T. Kim, *Big Data and Artificial Intelligence: New Challenges for Workplace Equality*, 57 U. LOUISVILLE L. REV. 313, 321 (2019) [hereinafter *Big Data and Artificial Intelligence*]; Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 860-61 (2017) [hereinafter *Data-Driven Discrimination*].

22. Cofone, *supra* note 11, at 151–58.

23. *Id.*

24. *See infra* Part III.

algorithmic discrimination. The algorithmic discrimination literature identifying these harms and challenges has so far brought useful legal approaches to an information problem (the problem of an information point being misused). This Article takes a different but complementary approach, and proposes that we must also establish an information solution to the informational problem that is algorithmic discrimination.

Part I explains the difference between knowledge-based, machine learning and deep learning algorithms, as well as the different ways in which these algorithms can lead to discriminatory outcomes: bias in the process, bias in the sample data, and social bias reflected in the data. Part II introduces the idea of an information approach to reduce discrimination. It explains how the law deploys such approach for human decision-makers and the specific challenges in applying this approach to algorithms. Part III develops how the information approach can be used with algorithmic decision-making. It separates between reducing, expanding, and transfer proxies, and it argues that we can and should deploy pre-processing techniques to either encode or shape training data. Part IV analyzes the doctrinal consequences of such a proposal: it overcomes the algorithmic fairness impossibility and the tension between disparate impact and disparate treatment in antidiscrimination law, it avoids the problem of algorithmic opacity to apply antidiscrimination measures, and it reaps the benefit of ex-ante regulation as a complement to ex-post liability.

I. ALGORITHMS CAN DISCRIMINATE

It is as dangerous as it is inaccurate to believe that, because algorithmic decision-making is computational, it cannot discriminate. Algorithms can be classified in knowledge-based and machine learning, which is the type this Article is centrally concerned with. These (machine learning) algorithms present the unfulfilled promise of unbiased decisions. The resulting discrimination can be classified along three categories. There can be bias among the people who create the algorithm that gets translated into the data-processing mechanism, there can be bias in the sample that is used by the algorithm, and there can be data that have inequality embedded in a way that leads to disparate impact. The first is a bias in the process, the second is a bias in the input (sample), and the third is a societal bias captured in representative data.

A. KNOWLEDGE-BASED, MACHINE LEARNING, AND DEEP LEARNING

Knowledge-based systems are traditional algorithms, which work when a computer scientist designs a list of decision rules for the algorithm to walk data points through, and can be visualized in the form of a flowchart. Machine

learning algorithms, on the other hand, extract those rules from their training data.[25]

Machine learning is a way of training an algorithm. While it became popular in the mid-2000s, its first definitions appeared much earlier. In 1959, Arthur Samuel notably called it an algorithm's "ability to learn without being explicitly programmed."[26] While knowledge-based algorithms are built on decision trees and detailed instructions indicating how to process data, machine learning algorithms are given large amounts of data with output variables for the algorithm to self-adjust. Instead of determining decision rules, human intervention is limited to selecting features for the training data and attaching labels to the output data.[27] There are many ways to do this, such as decision tree learning, reinforcement learning, clustering, and Bayesian networks.[28] Regardless of the mode in which the algorithm learns, the characteristic that is central to our purposes is that the model can learn decision rules.

Deep learning is a type of machine learning that became popular in 2012.[29] Deep learning uses a layered structure called an artificial neural network, which simulates a biological brain's neural network.[30] Like other forms of machine learning, deep learning algorithms collect training data, learn from it, and then apply what they learned to larger datasets to determine or predict something about reality. The difference is that, whereas machine learning requires feature selection, deep learning has automatic feature extraction. The algorithm engages in its own feature selection, adding layers of features that are used to map real-world data to a specific outcome.[31] That is, they adjust the weight given to each neuron and type of information.[32] This means that deep learning does not require a data scientist to intervene where its predictions are unsatisfactory—the neural network will make decisions about how its approach should change.[33] For example, self-driving cars use deep learning algorithms to recognize obstacles. The more that you "drive" your self-driving car, the better it will get at recognizing obstacles. While deep learning is not currently used in decision-making applications as are other machine learning models, it is by no means a

25. Karen Hao, *What is Machine Learning? We Drew You Another Flowchart,* MIT TECH. REV. (Nov. 17, 2018), https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/.

26. MARIETTE AWAD, RAHUL KHANNA, EFFICIENT LEARNING MACHINES: THEORIES, CONCEPTS, AND APPLICATIONS FOR ENGINEERS AND SYSTEM DESIGNERS 1 (2015); A. L. Samuel, *Some Studies in Machine Learning Using the Game of Checkers*, 3 IBM J. RES. DEV. 210 (1959).

27. Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2225 (2019).

28. FRY, *supra* note 2, at 56–59.

29. Dave Gershgorn, *The Data that Transformed AI Research—and Possibly the World*, QUARTZ (July 26, 2017), https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/.

30. *See* Lehr & Ohm, *supra* note 17, at 670.

31. IAN GOODFELLOW ET AL., DEEP LEARNING 8 (2016).

32. *Id.*

33. *Id.* ("Deep learning is a particular kind of machine learning that achieves great power and flexibility by representing the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones.").

stretch of the imagination to see it increasingly applied to decision-making in the near future.

B. THE UNFULFILLED PROMISE OF UNBIASED DECISION-MAKERS

Algorithmic decision-making is sometimes taken to imply that the prevalence of biases for discrimination decreases. While we humans are flawed and, even when well-meaning, host a wide array of implicit biases, algorithms are often presented as fairer and unbiased decision-making agents.[34]

However, in the last few years, piles of documented cases have appeared regarding decision-making processes in which algorithms also produce a discriminatory outcome—even assuming no discriminatory intent.[35] Examples of this exist in almost any area of decision-making, but perhaps the most prevalent are criminal procedure and employment.

The first example, as well-known as it is illustrative, pertains to criminal procedure. A few years ago, Northpointe (now called Equivant) developed a risk assessment algorithm called COMPAS to use as a recidivism indicator: the algorithm was designed to predict the likelihood of a person to re-offend within two years.[36] COMPAS is widely used to predict the likelihood that people who have been arrested will commit future crimes, and to determine parole. In a now classic article, ProPublica accused COMPAS of producing racially biased results, having almost twice as many false positives for black defendants than for white defendants and more frequent false negatives for white defendants than for black defendants.[37] In other words, the algorithm mistakenly identified as high risk twice as many black individuals than it did white individuals, and

34. There are many academic references that illustrate this position but perhaps the most interesting illustration is one of popular culture. In the movie, MONEYBALL (Columbia Pictures 2011), Billy Beane (Brad Pitt) was hired to help choose baseball players for a team. At the beginning of the movie, there is a scene in which he observes the previous coaches choose the next player based on factors such as whether they look confident, are good looking, or have a girlfriend. Beane condescendingly suggests, instead, to have an "automated" process based on the performance statistics of each player. This is presented in the movie, and outside of it, as a fair and impartial process—and more often than not, it is one.

35. Barocas & Selbst, *supra* note 21; Toon Calders & Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY: DATA MINING AND PROFILING IN LARGE DATABASES 43, 55–56 (Bart Custers et al. eds., 2013); Nizan Geslevich Packin & Yafit Lev-Aretz, *Learning Algorithms and Discrimination*, *in* RESEARCH HANDBOOK ON THE LAW OF ARTIFICIAL INTELLIGENCE (Woodrow Barfield & Ugo Pagallo eds., 2018); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 8–18 (2014); Tal Z. Zarsky, *An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics*, 14 I/S: J.L. & POL'Y INFO. SOC'Y 11, 15–22 (2017).

36. Tim Brennan et al., *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 CRIM. JUST. & BEHAV. 21, 22–24 (2009).

37. Julia Angwin et al., *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 (finding the false positives to be 23.5% for white defendants and 44.9% for black defendants, and finding the false negatives to be 47.7% for white defendants, and 28.0% for black defendants); Jeff Larsen et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

mistakenly identified as low risk more white individuals than it did black individuals.[38] Since then, dozens of papers have been written criticizing COMPAS.[39]

COMPAS is not the only algorithm to have a disparate impact when making risk assessment predictions. A different risk assessment algorithm, used at the federal level to make probation decisions, was also found to give a higher average score of post-conviction risk assessment to black individuals.[40] Even if the study may conclude that bias is unlikely because 66% of the racial difference was attributable to criminal history, and criminal history is not a proxy for race, criminal history does affect the relationship between race and future arrest; this is because, for the same criminal activity, black individuals are more likely to be arrested than white individuals.[41]

An illustrative employment example is a machine learning system that Amazon recently developed to rank job candidates. The system displayed a significant bias against female candidates, justifiably triggering public outcry.[42]

---

38. In addition, COMPAS was accused of being as accurate as a simple predictor based on prior count both on false positives and on false negatives—even when COMPAS and the prior count predictor may disagree. While the inaccuracy problem and the discrimination problem are distinct, the accusations of inaccuracy should raise eyebrows. *See* Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 153, 156 (2017).

39. Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017); Fiona Doherty, *Obey All Laws and Be Good: Probation and the Meaning of Recidivism*, 104 GEO. L.J. 291 (2016); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59 (2017); Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231 (2015); Hamilton, *supra* note 15; Kelly Hannah-Moffat, *Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates*, THEORETICAL CRIMINOLOGY, 2018; Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043 (2019); Richard F. Lowden, *Risk Assessment Algorithms: The Answer to an Inequitable Bail System?*, 19 N. C. J. L. & TECH. 221 (2018); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014); Danielle Kehl et al., *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, RESPONSIVE COMMUNITIES INITIATIVE, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y (July 2017), https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf.

40. Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 685 (2016).

41. *Id.* at 700.

42. James Cook, *Amazon Scraps "Sexist AI" Recruiting Tool that Showed Bias Against Women*, TELEGRAPH, (Oct. 10, 2018), https://www.telegraph.co.uk/technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/; Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS, (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G; Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, STAR ONLINE, (Oct. 10, 2018), https://www.thestar.com.my/tech/tech-news/2018/10/10/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women/; Roberto Iriondo, *Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women*, MEDIUM, (Oct. 11, 2018), https://medium.com/datadriveninvestor/amazon-scraps-secret-ai-recruiting-engine-that-showed-biases-against-women-995c505f5c6f; Cathy O'Neil, *Amazon's Gender-Biased Algorithm Is Not Alone*, BLOOMBERG, (Oct. 16, 2018), https://www.bloomberg.com/opinion/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone; James Vincent, *Amazon Reportedly Scraps Internal AI Recruiting Tool that Was Biased Against Women*, VERGE, (Oct. 10, 2018), https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report; Jordan Weissmann, *Amazon Created a Hiring Tool Using AI. It Immediately Started Discriminating Against Women.*, SLATE, (Oct. 10, 2018), https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html.

Because the algorithm was trained using Amazon's existing hiring data under the idea that Amazon's current employee choices are a good proxy for Amazon's desired employee choices, the algorithm reflected existing hiring practices. These practices, however, to the surprise of the algorithm's programmers, ended up being sexist.[43]

The Amazon algorithm illustrates a key concern of algorithmic discrimination: not only does automated decision-making mirror existing biases, but it has the potential to amplify them.[44] The types of societal biases held among human decision-makers are, to a large extent, consistent. Yet, human decision-makers exercise decisions on a case-by-case basis, implicating different levels of bias. Automated decision-making, on the other hand, brings perfect consistency across decisions. With this consistency, it brings the potential to discriminate systemically.[45] In such a way, existing patterns of discrimination embedded in machine learning models leads them to not only perpetuate, but directly contribute to, further marginalization.[46]

Many examples of algorithmic bias seem to indicate that algorithmic bias is most likely to affect disadvantaged populations: those who are more likely to find themselves asking for parole or who are more likely to apply for the type of jobs that use an algorithm to screen candidates.[47] While it is true that disadvantaged populations are disproportionately affected, it is not true that they are the only ones affected. Several examples illustrate this, most commonly in allocating healthcare, particularly by determining health insurance quotas,[48] and financial instruments, particularly through credit scores.[49] The Amazon

43. The algorithm, for example, panelized resumes that contained explicitly gendered words such as "woman," and candidates who went to colleges that were identified as all-women institutions. *See id*.

44. *See supra* Mayson, note 27, at 2251 ("[P]rediction functions like a mirror. The premise of prediction is that, absent intervention, history will repeat itself. So what prediction does is identify patterns in past data and offer them as projections about future events."); *Big Data and Artificial Intelligence*, *supra* note 21, at 320 ("If the employer's prior hiring practices excluded certain groups—for example, the hypothetical Tech Co., which hired very few women as computer programmers in the past—the algorithm will simply reproduce the previously existing biases. . . . [t]he selection tool might operate to exclude racial or ethnic minorities.").

45. Indrė Žliobaitė, A Survey on Measuring Indirect Discrimination in Machine Learning 4 (Oct. 2015), (unpublished manuscript), http://arxiv.org/abs/1511.00148 ("While human decision makers may make biased decisions on case by case basis, rules produced by algorithms are applied consistently, and may discriminate more systematically and at a larger scale."); *Big Data and Artificial Intelligence*, *supra* note 21, at 322 (discussing the making of decisions on a large scale, affecting whole groups rather than individual cases).

46. Barocas & Selbst, *supra* note 21, at 677–93; *Data-Driven Discrimination*, *supra* note 21, at 883–92; *Big Data and Artificial Intelligence*, *supra* note 21, at 321–22 ("If an algorithm erroneously predicts that I am pregnant and sends me coupons for diapers, I can simply ignore them. If, however, it predicts—erroneously—that I will not be a good employee and I am denied a job as a result, it has created a much more significant problem for me. And if an algorithm not only makes an erroneous prediction about an individual worker, but makes predictions across cases or populations in a way that is *systematically* wrong or biased—that raises much broader social concerns."); NOBLE, *supra* note 10.

47. *See, e.g.*, EUBANKS, *supra* note 4. *See* Latanya Sweeney, *Discrimination in Online Ad Delivery*, 11 ACM QUEUE, Apr. 2, 2013, at 1 (showing that arrest records advertisements are more likely to appear on searches for black-sounding names).

48. PASQUALE, *supra* note 1.

49. FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? UNDERSTANDING THE ISSUES 12 (2016), https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-

algorithm example, moreover, illustrates that the use of A.I. is not restricted to low-income employment.[50] The everyday life of every American citizen is affected by automated decision-making, whether they know it or not.

The promise of fair and unbiased algorithmic deciders has not delivered. It has failed to deliver not because algorithms are not useful to make decisions and we must get rid of them. Rather, it has failed to deliver because algorithmic discrimination is a discrimination problem among humans that, because it introduces a different type of interaction, must be regulated differently.[51]

C. BIASED PROCESS

The most evident type of algorithmic bias is a bias in the way in which an algorithm processes information: a bias in the model itself, or a classification bias. A biased process is the type of bias that traces most clearly to traditional antidiscrimination law, as treating two groups differently maps onto disparate treatment.[52]

We sometimes believe that we are interacting with an A.I. For example, when we have our credit score calculated for loans or credit card applications, our car speeds monitored for fines, or our risk determined for airport security checks. But instead, we are interacting, through the algorithm, with the humans that design and apply it.[53] Algorithms are not a new type of agent, but a new way in which we interact with other people. This insight has crucial implications for understanding discrimination. As Nick Seaver puts it, "[w]hile some proponents of algorithms—or machine learning, or artificial intelligence, or whatever complexly responsive software is called by the time you read this—may claim that their systems are autonomous, there are [people] everywhere, tweaking and tuning, repairing and refactoring."[54]

Realizing this fact is highly consequential. It means that, when analyzing how interactions with algorithms are structured, we can build on our knowledge

---

report; Tracy Alloway, *Big Data: Credit Where Credit's Due*, FIN. TIMES (Feb. 4, 2015), https://www.ft.com/content/7933792e-a2e6-11e4-9c06-00144feab7de.

50. *See* Amit Datta et al., *Automated Experiments on Ad Privacy Settings*, 1 PROC. PRIVACY ENHANCING TECH. 92, 95, 105 (2015) (showing that advertisements for high-income jobs are shown to men more than they are to women).

51. Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 124–28 (2014) (arguing that the predictive nature of big data changes the assumptions of current legislation and a right to procedural data due process is needed); Pauline T. Kim & Sharion Scott, *Discrimination in Online Employment Recruiting*, 63 ST. LOUIS U. L.J. (forthcoming 2019) (questioning whether current law of Title VII and the Age Discrimination in Employment Act are able to cover all forms of targeted recruitment given their outdated archetype of recruitment models); Mark A. Lemley & Bryan Casey, *Remedies for Robots* 104–07 (Stan. Law & Econ. Olin, Working Paper No. 523, 2019) (arguing that artificial intelligence and robots complicate the question of the appropriate remedy).

52. Barocas & Selbst, *supra* note 21, at 694–98; *see also infra* Subpart IV.A.

53. Jack M. Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1221 (2017); Zachary C. Lipton, *The Mythos of Model Interpretability*, ACM QUEUE, July 17, 2018, at 2–6.

54. Nick Seaver, *What Should an Anthropology of Algorithms Do?*, 33 CULTURAL ANTHROPOLOGY 375, 378 (2018).

of interactions with other humans.[55] When used to make decisions, algorithms, simply put, are tools for a person to use a large dataset to predict the desired output.[56]

Algorithms are never entirely autonomous.[57] Any decision-making algorithm requires a human to determine the desired output under a conditional probability ("given input X produce this output Y").[58] Under the examples mentioned above, supervised learning would consist of a person ordering "here is an email, output the probability of this email being spam" or "here is a potential tenant's data, output the probability of this tenant defaulting on payment." The same can be said under unsupervised learning, given that humans must choose the input data to develop (unlabeled) features and the output variable that serves as a proxy for the desired characteristic.

While bias in machine learning algorithms is often a data bias, humans must frame the problem and make a choice about what the algorithm should predict before any data are processed.[59] For an algorithm to be used, "recidivism" or "employability" must translate into something measurable.[60] Moreover, after determining the prediction's goal, humans must also decide which attributes they want the algorithm to consider in order for the algorithm to determine whether and how much each of the attributes is predictive of the stated goal.[61]

---

55. *See* Cofone, *supra* note 14 (arguing that the legal problems posed by A.I. are problems among humans intermediated by A.I. and, based on that idea, proposing different analogies to address those problems).

56. AGRAWAL ET AL., *supra* note 4; *see also* Bryan Pfaffenberger, *Fetishised Objects and Humanised Nature: Towards an Anthropology of Technology*, 23 MAN 236, 241 (1988) ("Technology is not an independent, non-social variable that has an 'impact' on society or culture. On the contrary, any technology *is* a set of social behaviours and a system of meanings. To restate the point: when we examine the 'impact' of technology on society, we are talking about the impact of one kind of social behaviour on another."); Seaver, *supra* note 54, at 382 ("[A]nthropology's greatest contribution to contemporary critical discourse about algorithms may be the corrosive potential of anthropological attention itself, which promises to break down the apparent solidity and coherence of the algorithm.").

57. Seaver, *supra* note 54, at 378 ("[P]ress on any algorithmic decision and you will find many human ones: people like Brad or his manager deciding that a certain error threshold is acceptable, that one data source should be used over another or that a good recommendation means this, not that. These systems are, as a former head of recommendation at Spotify put it, 'human all the way down' (citation omitted). There is no such thing as an algorithmic decision; there are only ways of seeing decisions as algorithmic.").

58. This claim is often made for machine learning more generally. *See* Lipton, *supra* note 53. *See* Seaver, *supra* note 54, at 378 ("While discourses about algorithms sometimes describe them as 'unsupervised,' working without a human in the loop, in practice there are no unsupervised algorithms. If you cannot see a human in the loop, you just need to look for a bigger loop."). While there are some examples of unsupervised learning, such as clustering and dimensionality reduction, none of them pertain predictive algorithms used for decision-making.

59. Barocas & Selbst, *supra* note 21, at 677–84 (explaining that programmers may introduce bias into the algorithm when choosing training data or choosing a target variable); *see also* Balkin, *supra* note 53, at 1223 ("But in most cases, the problem isn't the robots; it's the humans.").

60. Samir Passi & Solon Barocas, *Problem Formulation and Fairness*, *in* PROC. OF THE CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 39–48 (2019), http://doi.acm.org/10.1145/3287560.3287567 (explaining how bias is generated in the process of identifying goals).

61. Citron & Pasquale, *supra* note 35, at 14 ("Software engineers construct the datasets mined by scoring systems; they define the parameters of data-mining analyses; they create the clusters, links, and decision trees applied; they generate the predictive models applied. The biases and values of system developers and software

The biases of humans that program and apply the algorithm can translate into the algorithm, and sometimes stereotypes and negative associations can be codified in and amplified by the algorithm.[62] Algorithms, after all, are built, trained, and implemented by people that, like everyone else, have prior beliefs and goals determined by social factors.[63]

Researchers have known for a long time that people's prior beliefs translate into the algorithms that they generate.[64] Common behavioral biases that could turn into a model's assumptions are, for example, reporting bias, selection bias, and availability bias. These could be translated, for example, into faulty labeling, faulty personalization that leads to filter bubbles or tunnel vision, incorrectly assuming causation, or one side of the matching incorrectly given an advantage over the other side. Facial recognition machine learning software, for example, have been shown to be affected by the demographics of the people who design them.[65] This issue is exacerbated if programmers are disproportionately male, white, and heterosexual.[66]

A key reported challenge to applying disparate treatment to a biased algorithmic process is developing a theory of intent.[67] But the relevant intent

programmers are embedded into each and every step of development."); Seaver, *supra* note 54, at 378-79 ("Take, for instance, the case of the programmer Jacky Alciné (2015), who discovered that Google Photos had automatically tagged a picture of him and a friend as containing 'gorillas.' . . . By the next day, the objectionable tag was gone. If we grant algorithms autonomy, treating them as though they are external to human culture, then we cannot make any sense of this story—why it happened in the first place, how it was resolved, and why similar problems might happen again. Tales of autonomous algorithms cannot explain why the system works differently now than it did then. To explain what happened, we need to talk about the makeup of technical teams, the social processes by which those teams define and discover problems, how they identify acceptable solutions, and their culturally situated interpretive processes.").

62. Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 553–58 (2018) (discussing the potential applicability of Title VII's stereotype theory of liability to algorithms).

63. NOBLE, *supra* note 10, at 1–2 (pointing out that humans are behind the algorithm); *see also* Barocas & Selbst, *supra* note 21, at 677–84 (explaining that programmers may introduce bias into the algorithm when choosing training data or choosing a target variable).

64. Langdon Winner, *Do Artifacts Have Politics?*, 109 DAEDALUS 121, 128–35 (1980); *see also* WHITTAKER, ET AL., *supra* note 18, at 38–40.

65. P. Jonathon Phillips et al., *An Other-Race Effect for Face Recognition Algorithms*, 8 ACM TRANSACTIONS APPLIED PERCEPTION, Jan. 2011, at 14:1, 14:7; Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. OF MACHINE LEARNING RES. 1 (2018), 1–2, http://proceedings.mlr.press/v81/buolamwini18a.html [https://perma.cc/HRR9-69HX].

66. Kate Crawford, Opinion, *Artificial Intelligence's White Guy Problem*, N.Y. TIMES (June 25, 2016), https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html. This over-representation appears to be largest in terms of gender. *See* Yoan Mantha & Simon Hudson, *Estimating the Gender Ratio of AI Researchers Around the World*, MEDIUM (Aug. 17, 2018), https://medium.com/element-ai-research-lab/estimating-the-gender-ratio-of-ai-researchers-around-the-world-81d2b8dbe9c3 (estimating a cross-country average of 88% of AI researchers being male, the number being 86.57% in the US); Tom Simonite, *AI Is the Future—But Where Are the Women?*, WIRED (Aug. 17, 2018), https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/ (estimating that only 10-15% of A.I. research staff in Facebook and Google are women, and only 12% of researchers at major A.I. conferences are women).

67. Sam Corbett-Davies & Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning 4 (Aug. 14, 2018) (unpublished manuscript), http://arxiv.org/abs/1808.00023 (arguing that law's reliance on intent is unsuited to deal with algorithmic systems).

comes from the humans involved in programming, training, and applying the algorithm.[68] Even if there is no direct human intent over the outcome, there is always human involvement in how the decision is made. Also in machine learning algorithms, through data-gathering, feature selection, and choice of target variable, there is human involvement in any algorithmic model—trained or untrained. People must select what data can predict the relevant features, what features are important enough for the algorithm to consider in order to determine the output and, perhaps most importantly, people must choose the target variable that serves as a proxy for the desired output.[69]

In sum, there is always, at some level, a human decision-maker that impacts the process. Biases in an algorithmic process often exist because human biases were translated into the system. These biases are inevitable, but they can be reduced when the environment forces them to be made explicit.

D. BIASED SAMPLE DATA

An algorithm can only be as good as the data that it is fed. If an algorithm is mining in a section of the dataset that, for any reason, is unrepresentative of the population, it will produce a non-representative output.

Bias in the data has been explored by prior research, in particular by Solon Barocas and Andrew Selbst.[70] The problem of biased data exists for both individual records in a dataset and for the dataset as a whole. Individual records may suffer from quality problems due to partial or even incorrect data. The entire dataset might have quality problems at higher rates for an entire protected class compared to others or might be unrepresentative of the general population.[71]

These variations in the quality and representativeness of data may correlate with class membership and, in turn, negatively impact historically disadvantaged groups when used to make decisions about members of these groups.[72] Since datasets involving historically disadvantaged groups can suffer data quality problems for a variety of reasons, such as underrepresentation, the use of these datasets can increase discriminatory outcomes.[73]

This is a more sophisticated version of a standard statistics problem that is usually solved by obtaining more data as, according to the law of large numbers, when the sample is large enough it will resemble the population.[74] In this

68. *Data-Driven Discrimination*, *supra* note 21, at 884.

69. Because there are theories of human behavior behind every algorithm, disparate treatment is applicable. *Cf. id* at 890–911 (arguing a different position: that disparate impact is, most of the time, inapplicable to machine learning algorithms, but Title VII includes a prohibition of classification bias that is consequence-based, but distinct from antisubordination).

70. Barocas & Selbst, *supra* note 21, at 684–86.

71. *Id.* at 684–87.

72. *Id.*

73. Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55, 61 (2013); Kate Crawford, *Think Again: Big Data*, FOREIGN POL'Y (May 10, 2013), http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data.

74. STEWART J. ANDERSON, BIOSTATISTICS: A COMPUTING APPROACH 60 (2012). This is, of course, as long as the sample is large enough that is unbiased with respect to the population.

context, however, obtaining more data often fails to solve the problem, either because the algorithm uses an already existing biased database or because it works through a machine learning process that continues to obtain biased samples. [75] Oftentimes, the data fed to algorithms suffer from a self-selection problem.

To see how this may work in the analogue world, imagine that the New York Police Department decided to engage in predictive policing based solely on prior arrest numbers. If it had more of its police force in the Bronx than in other boroughs, then it would be likely to make more arrests there than anywhere else. This would lead to more comparative police presence in the Bronx, leading to more comparative arrests, and so on and so forth, independent of the actual crime rates.[76]

This self-selection issue is more severe in algorithmic decision-making because any bias risks becoming systematic.[77] This leads to an unhelpful feedback loop problem whereby algorithms find correlations in a biased dataset and then predict outcomes without taking into account the fact that bias tainted the training data. In successive rounds of analysis and prediction, the bias that began in the dataset then determines future outcomes. These predictions are then fed back into the algorithm, creating a vicious circle.[78]

Predictive policing algorithms such as HunchLab and PredPol often work in a similar way: the algorithm detects crime patterns in a city, it predicts the likelihood of future crime per geographic area based on such patterns, and the police patrol not randomly, but according to such predictions.[79] Because the police will invariably find more crime where it is looking for crime than where it is not, and the new data on arrests will be fed to the algorithm, this will lead

75. Citron & Pasquale, *supra* note 35, at 5 ("Although software engineers initially identify the correlations and inferences programmed into algorithms, Big Data promises to eliminate the human 'middleman' at some point in the process. Once data-mining programs have a range of correlations and inferences, they use them to project new forms of learning." (footnote omitted)).

76. Rashida Richardson et al., *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U L. REV. 193, 193 (2019) (discussing how systemic data manipulation, falsifying police reports, unlawful use of force, planted evidence, and unconstitutional searches lead to systemic biases in datasets used for predictive policing); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 118–19 (2017) (showing how data-driven policing risks perpetuating discrimination and proposing algorithmic impact statements to inform about potential disparate impact).

77. *See, e.g.*, Tolga Bolukbasi et al., Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings 3 (July 21, 2016) (unpublished manuscript), https://arxiv.org/pdf/1607.06520.pdf (discussing how word embeddings can amplify gender bias)*; see also* O'NEIL, *supra* note 8, at 200.

78. Lemley & Casey, *supra* note 51, at 57 (discussing the difficulties in granting effective remedies for discriminatory algorithmic decision-making due to this kind of feedback loop).

79. Aaron Shapiro, *Reform Predictive Policing*, 541 NATURE 458, 458 (2017) ("Geospatial modelling generates risk profiles for locations. Jurisdictions are divided into grid cells (each typically around 50 square metres), and algorithms that have been trained using crime and environmental data predict where and when officers should patrol to detect or deter crime."); *see also New Model Police*, ECONOMIST (June 7, 2007), https://www.economist.com/united-states/2007/06/07/new-model-police. Like COMPAS, HunchLab (as well its main competitors such as PredPol) is a proprietary algorithm, so it is not possible to know exactly how it makes predictions.

to a feedback loop problem.[80] A dataset with a feedback loop problem such as this one will be likely to adversely affect vulnerable minorities.[81]

E. DATA THAT REFLECT A BIASED SOCIETY

On January 22, reporter Ryan Saavedra attempted to ridicule Rep. Alexandria Ocasio-Cortez by sharing a video of her in which, in his words, she "claims that algorithms, which are driven by math, are racist."[82] In the over 7,000 replies to the tweet (more than what tweets about algorithms usually receive), several researchers who work on algorithmic decision-makers, along with members of the public who linked to their articles, attempted to explain that she was correct. A few years ago, his statement would have gone unnoticed, but the public and policy attention to the ways in which algorithms can discriminate is on the rise.[83]

A machine learning algorithm's training data not only can be biased or incomplete, as discussed in the last Subpart, but can also reflect prior discrimination.[84] If the training data are biased, as was the Amazon hiring algorithm's data—which reflected the existing male dominance in the tech industry—the outcome will be biased.[85]

An algorithmic process can produce a disparate impact even when trained with representative data.[86] The difference between the biased data from the last Subpart and this type of bias is that, here, the data are representative of the

80. Kristian Lum & William Isaac, *To Predict and Serve?*, SIGNIFICANCE Oct. 2016, at 16 (running a simulation on PredPol that seems to indicate that the algorithm amplifes database biases due to feedback loops); Danielle Ensign et al., Runaway Feedback Loops in Predictive Policing 1–2 (Dec. 22, 2017) (unpublished manuscript), http://arxiv.org/abs/1706.09847 (exploring when feedback loops occur).

81. *See, e.g.*, Nathan Munn, *Police in Canada Are Tracking People's 'Negative' Behavior* in *a 'Risk' Database*, VICE (Feb. 27, 2019), https://www.vice.com/en_us/article/kzdp5v/police-in-canada-are-tracking-peoples-negative-behavior-in-a-risk-database ("[T]he algorithm underpinning PredPol, one of the most widely used predictive policing technologies, is fundamentally flawed in a way that can contribute to over-policing, particularly for marginalized communities."). The Chicago Police Department's Strategic Subject List is another example of the same problem. The Strategic Subject List's algorithm predicts the likelihood of a person being involved in gun violence in the future, either as a victim or perpetrator. However, because the list was allegedly used by the police department as an informal suspect list for crimes involving gun violence, it was shown to be predictive not of involvement in future gun violence but of probability of being arrested in the future. *See* Jessica Saunders et al., *Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. EXPERIMENTAL CRIMINOLOGY 347, 363–64 (2016).

82. Ryan Saavedra (@RealSaavedra), TWITTER (Jan. 22, 2019, 12:27 AM), https://twitter.com/RealSaavedra/status/1087627739861897216.

83. *See, e.g.*, John Burn-Murdoch, *The Problem with Algorithms: Magnifying Misbehaviour*, GUARDIAN (Aug. 14, 2013), https://www.theguardian.com/news/datablog/2013/aug/14/problem-with-algorithms-magnifying-misbehaviour; Claire Cain Miller, *When Algorithms Discriminate*, N.Y. TIMES (July 9, 2015), https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html.

84. *See generally* Calders & Žliobaitė, *supra* note 35.

85. Cook, *supra* note 42; Dastin, *Amazon scraps secret AI*, *supra* note 42; Dastin, *Amazon scraps "sexist AI*," *supra* note 42; Iriondo, *supra* note 42.

86. Barocas & Selbst, *supra* note 21, at 691–92.

population, but this representative data still produce a disparate impact outcome because of embedded social inequalities.[87]

To understand why a model trained with representative data may lead to disparate impact discrimination, we can refer to the idea of statistical discrimination.[88] Imagine an employer-deployed algorithm that cannot observe each worker's skill level, drawn from a normal, bell-shaped skill distribution. The algorithm, however, can observe two things. First, it can observe their group identity. We can define such identity in any way, such as *P* for purple hair and *G* for green hair. Second, it can observe a noisy or imprecise signal about each person's productivity. Under the model, the question of statistical discrimination is the question of why two workers with the same productivity signal, but from different groups, are treated differently.[89]

This differential treatment would take place under two scenarios: stereotyping and differential observability. Stereotyping takes place when all signals are equally informative of each individual's productivity, but one group, *P*, has a lower average human capital investment, potentially leading to lower average skill. Because the algorithm will take both group membership and the signal to be informative of each individual's expected productivity, it will consider an employee from *P* to have a lower *expected* productivity than an employee from *G* that has the same signal.[90] Therefore, *P* workers will receive a lower salary under the same signal, or they will be offered fewer jobs.

Differential observability takes place when the skill distributions are identical in both groups, but the signals for *P* workers' skills are less informative than those of *G* workers. This will lead the algorithm to consider the expected productivity of a *P* worker with any signal to be closer to that of the population average than the expected productivity of a *G* worker with the same (although less noisy) signal because the signals for *G* and for *P* will be given different weights. This, in turn, will lead highly qualified *P* workers to receive a lower salary than their *G* equivalents, and low qualified *P* workers to receive a higher salary than their *G* equivalents.[91]

---

87. Aylin Caliskan et al., *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCI. 183, 183 (2017). *See generally* NOBLE, *supra* note 10; DANIEL ROSENBERG, ET AL., "RAW DATA" IS AN OXYMORON (Lisa Gitelman ed., 2013).

88. Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659, 661 (1972); Kenneth. J. Arrow, *The Theory of Discrimination*, *in* DISCRIMINATION IN LABOR MARKETS (Orley Ashenfelter & Albert Rees eds., 1973).

89. *See* Shelly Lundberg & Richard Startz, *On the Persistence of Racial Inequality*, 16 J. LAB. ECON. 292, 292–95 (1998) (introducing a model showing that statistical discrimination in competitive markets and without differences in average human capital introduces inefficiencies in the system; due to statistical discrimination, minorities face lower incentives to invest in human capital, community social capital is lowered, and they develop lower levels of productivity).

90. As noted above, algorithms can only make predictions about the likelihood of a future event, and cannot predict any event: it can determine the likelihood of someone to be a productive employee, but not whether they will be one.

91. Hanming Fang & Andrea Moro, *Theories of Statistical Discrimination and Affirmative Action: A Survey*, *in* 1A HANDBOOK OF SOCIAL ECONOMICS 133, 137–40 (Jess Benhabib et al. eds., 2011).

While it is difficult to infer the type of statistical inference without looking at the data and how they were processed, one could imagine the COMPAS algorithm falling in stereotyping and the Amazon hiring algorithm falling in differential observability. COMPAS, as detailed below,[92] was working with a dataset where black individuals had more arrests on average than white individuals due to embedded societal inequalities, and this fact led to its discriminatory outcome. Amazon's algorithm had a database that was mostly male and could therefore learn about expected productivity for male candidates better than it did for female candidates,[93] resulting in a discriminatory outcome of its own. If the data sample is representative of reality, it will also reflect the existing prejudices that exist.

Some historically disadvantaged groups (including along the dimensions of race and gender) are protected from these types of statistical discrimination under Title VII of the Civil Rights Act,[94] even when the generalizations are true.[95] There is agreement that statistical discrimination should be eradicated, but we lack ways of doing so effectively. As will be explored in the rest of this Article, antidiscriminatory information rules help inform efforts to address algorithmic discrimination.

II. AN INFORMATION APPROACH TO ALGORITHMIC DISCRIMINATION

Because algorithmic decision-making is still human decision-making mediated by algorithms, we can build on our knowledge about how humans operate to evaluate how to regulate algorithms optimally.[96] This Part explains when blocking information can prevent discrimination in human decision-making and why these privacy rules blocking information are especially suited for algorithms. Antidiscriminatory information rules are especially warranted when it comes to algorithms because algorithmic bias is based directly on the information that algorithms are fed. It then explores the key challenge to apply this framework to algorithmic decision-making: data that are proxies for protected categories. The framework shows that the relationship between privacy and discrimination depends on the types of proxies that decision-makers rely on.

A. PRIVACY RULES THAT PREVENT DISCRIMINATION

Law often blocks personal information from human decision-makers in many areas (including in employment, health care, and banking) to prevent discrimination. In prior research, I proposed a framework for determining when personal information should flow and when it should not in order to prevent

92. *See infra* Subpart IV.C.
93. *See supra* Subpart I.A for an explanation of how machine learning algorithms learn.
94. 42 U.S.C. 2000e (2012).
95. City of L.A. Dep't of Water & Power v. Manhart, 435 U.S. 702, 716–17 (1978).
96. *See infra* Subpart II.A.

discrimination by humans. The success of these measures depends on what types of proxies exist for the information blocked.[97]

While traditional approaches to discrimination respond to the way a decision-maker *used* information, a preventive approach bars decision-makers from even acquiring the information. This prevents the decision-maker from taking an action in violation of antidiscrimination law.[98] It is therefore key to determine when personal information should flow in order to design privacy rules that prevent discrimination effectively. [99]

There are three conditions under which privacy rules reduce discrimination in a way that giving decision-makers more information does not.[100] These scenarios help determine when an antidiscriminatory information rule is warranted.

First, blocking certain data points can be useful when people do not update prior beliefs cleanly[101]—perhaps due to representativeness heuristics and confirmation biases.[102] People have limited time and attention to receive information, even when the information is unlimited.[103] Moreover, once people form a belief, most do not update their prior beliefs as cleanly as an ideal rational actor would when new information arrives.[104] The importance of this effect will be different depending on context and on the informational demands of the decision. In algorithmic decision-making, this translates into bias in the process.[105]

Second, blocking future information is useful when information samples are expected to be skewed and only a non-infinite amount of information can be gathered.[106] In algorithmic decision-making, this translates into bias in the sample data.[107]

---

97. Cofone, *supra* note 11.

98. Lisa Austin, *Privacy and the Question of Technology*, 22 L. & PHILOS. 119, 144 (2003); *Preempting Discrimination*, *supra* note 12, at 440–41; *Protecting Privacy*, *supra* note 12, at 2100.

99. Cofone, *supra* note 11, at 40–41.

100. *Id*. at 13–15. This differs from the standard economic approach to the relationship between privacy and discrimination. *See* Lior Jacob Strahilevitz, *Privacy Versus Antidiscrimination*, 75 U. CHI. L. REV. 363, 364 (2008) (presenting such approach); *see also infra* Subpart II.B.

101. Cofone, *supra* note 11, at 150.

102. *See generally* David M. Grether, *Bayes Rule as a Descriptive Model: The Representativeness Heuristic*, 95 Q. J. ECON. 537 (1980); Raymond S. Nickerson, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, 2 REV. GEN. PSYCHOL. 175 (1998).

103. JEFFREY ROSEN, THE UNWANTED GAZE: THE DESTRUCTION OF PRIVACY IN AMERICA 9–10 (2001); Lawrence Lessig, *Privacy and Attention Span*, 89 GEO. L. J. 2063, 2063–64 (2001).

104. Jennifer L. Eberhardt et al., *Seeing Black: Race, Crime, and Visual Processing*, 87 J. PERSONALITY & SOC. PSYCHOL. 876, 877 (2004); Brian A. Nosek et al., *Pervasiveness and Correlates of Implicit Attitudes and Stereotypes*, 18 EUR. REV. SOC. PSYCHOL. 36 (2007).

105. *See infra* Subpart II.B; s*ee also* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1129 (2018).

106. Cofone, *supra* note 11, at 149–50. Therefore, the law of large numbers would not solve the biased data problem.

107. *See infra* Subpart II.C; *see also* Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 1995–97 (2017).

Third, blocking information points can be a useful preventive mechanism when, due to larger societal values, the law simply does not want an information point to be considered even if it can be useful to a decision-maker.[108] This is the case of intentional statistical discrimination, where a decision-maker purposely employs a heuristic for information cost saving, but that heuristic discriminates against a protected class. In algorithmic decision-making, it translates into data that reflect a biased society.[109]

These rules can offer short-term protection from discrimination by blocking information that the law deems harmful in a decision-making process. When decision-makers' samples are skewed, when they have processing errors such as behavioral biases, or when discrimination is intentional, blocking information might be more effective at preventing discrimination than allowing it to flow. In this way, privacy rules can aid antidiscriminatory efforts.[110]

For all three scenarios, the key element in establishing an effective antidiscriminatory information rule is noting that, when information is blocked to prevent a discriminatory decision, oftentimes decision-makers use other information as proxies for such blocked information.[111] For example, decision-makers could use zip code as a proxy for race, or height as a proxy for gender. To be effective, the privacy rule must address those proxies as well.

A crucial task for preventing discrimination through privacy rules, therefore, is identifying and blocking data points that are proxies for categories that the law protects. The types of proxies determine under which conditions blocking an information flow will successfully tackle discrimination. In algorithmic decision-making, we have little experience with different types of proxies. We hear of cases in which blocking information worsened discrimination but, as with human decision-makers, we do not yet know when it does work.[112]

B. WHY PRIVACY RULES ARE ESPECIALLY SUITED FOR ALGORITHMS

Oftentimes in economics, discrimination is described as a problem of not having enough information about others.[113] According to this account, having insufficient information leads people to resort to heuristics to judge others, which can easily result in false opinions. These false opinions, in turn, are attributed to everyone who falls under the heuristic, resulting in beliefs that could be racist, sexist, or homophobic.[114] Recall, for example, the two situations of statistical discrimination: stereotyping and differential observability. The problem would not exist if the employer had perfectly informative productivity

108. Cofone, *supra* note 11, at 151.
109. *See infra* Subpart II.D.
110. Cofone, *supra* note 11, at 149–51.
111. *See infra* Subpart III.A.
112. *See infra* Part III.
113. *See, e.g.*, Strahilevitz, *supra* note 100, at 364, 380.
114. *See id* at 364.

signals.[115] In other words, describing discrimination as a problem of not having enough information about others assumes that discrimination is of a statistical nature, rather than based on psychological biases.

If discrimination were always statistical, making information about oneself more available would avoid the need for such heuristics and therefore reduce discrimination, while having more privacy would worsen it.[116] Representing this view, Lior Strahilevitz has argued that "by increasing the availability of information about individuals, we can reduce decision-makers' reliance on information about groups"[117] and that, therefore, "there is often an essential conflict between information privacy protections and antidiscrimination principles, such that reducing privacy protections will reduce the prevalence of distasteful statistical discrimination."[118] The underlying idea of this economic view is that society should tolerate statistical discrimination because the only way to dispense with it is by providing decision-makers with more information in contexts where there may be a normative reason for which they should not have access to it.[119]

This conclusion about the relationship between privacy and discrimination as being in tension operates under a strict set of assumptions: that humans are rational in their decision-making and they update biases cleanly.[120] These assumptions rarely apply to human decision-making, which is not always statistical and rarely involves a fully rational updating of beliefs.[121] However, the same cannot be said of algorithmic decision-making.

The assumptions of statistical discrimination and rational Bayesian updating of beliefs do apply in a straightforward way to algorithmic decision-making. Algorithms predict output variables based on data inputs. They are not irrational, they do not have prejudices, and they do not have limited time and attention. They operate precisely like the rational decision-maker of a standard economic model.

This makes the economic arguments about privacy and discrimination more relevant than before. Blocking information from an algorithmic decision-maker can only lead the system to perform more statistical inferences that are

115. *See supra* Subpart I.E.

116. *See* Shawn D. Bushway, *Labor Market Effects of Permitting Employer Access to Criminal History Records*, 20 J. CONTEMP. CRIM. JUST. 276, 288–89 (2004).

117. Strahilevitz, *supra* note 100, at 364.

118. *Id*; *see also* Lior Jacob Strahilevitz, *Reputation Nation: Law in an Era of Ubiquitous Personal Information*, 102 NW. U. L. REV. 1667, 1682–88 (2008).

119. Strahilevitz, *supra* note 118, at 1723–36. Note that this suggestion cannot eliminate statistical discrimination—it can only make it more targeted. Providing more information may be effective as an individual strategy when information is not blocked for the decision-maker, but these measures, while effective for some individuals, cannot eliminate statistical discrimination. *See, e.g.*, Joni Hersch & Jennifer Bennett Shinall, *Something to Talk About: Information Exchange under Employment Law*, 165 U. PA. L. REV. 49, 86–87 (2016) (finding that concealing family information lowers female applicants' hiring prospects).

120. *See* Strahilevitz, *supra* note 100, at 364; *see also* Strahilevitz, *supra* note 118, at 1675; *see infra* Subart III.A.

121. Cofone, *supra* note 11, at 150.

potentially discriminatory. These inferences will differ depending on the availability of proxies.

The limited applicability of the information approach to human decision-making stems precisely from the source of people's biases: privacy rules do not solve human biases. But algorithms' biases are based on the information that we feed them. Their source of bias is the input. Unlike humans, algorithms cleanly update their decisions when the input is changed. Therefore, most of the reasons for the limited applicability of this method to humans do not exist for algorithms.

Algorithms, in other words, are the rational decision-makers that humans can never be. The next Subpart will explore what this idea means for antidiscriminatory information rules in terms of the importance and impact of quality data.

C. IT IS ALL IN THE DATA

The discussion about algorithms above has illustrated that an algorithmic decision-making process can only be as good as the data that it uses. This insight about bias prevention is the reason why the conversation about algorithmic discrimination is a conversation about disparate impact. Without a focus on disparate impact, because the problem lies in the data and not in classifiers, the legal discussion on algorithmic discrimination has no grounding.[122]

This takeaway shows that algorithmic bias precedes the algorithm, because the bias exists in the data that are fed to the algorithm.[123] The data used to construct a machine learning system determine how the system interprets the world over which it operates to make predictions.[124] It is all in the data sample (and the humans that make it). From this point of view, it is unsurprising that regulating such data is a promising approach to aid antidiscriminatory efforts.

Because algorithmic bias precedes the algorithm, and because the bias exists in the data that are fed to the algorithm, antidiscriminatory information rules are, or should be, more useful to address algorithmic discrimination than they are for humans. Unlike humans, algorithms can block individual data-points in the decision-making process. While it can be difficult to instruct a human decision-maker to disregard a visible fact, it is more feasible, even if not always simple, to code an algorithm to do so. One can program an algorithm to control for any variable. For example, comparing a hiring outcome when including applicants' religious information and when blocking it. In other words,

---

122. *See generally* Barocas & Selbst, *supra* note 21, at 671–74; *Big Data and Artificial Intelligence*, *supra* note 21.

123. Batya Friedman & Helen Nissenbaum, *Discerning Bias in Computer Systems*, *in* INTERACT '93 AND CHI '93 CONF. COMPANION ON HUMAN FACTORS IN COMPUTING SYSTEMS 141 (Stacey Ashlund et al. eds., 1993), (separating preexisting, technical, and emergent biases, and arguing that freedom from bias should part of the normative criteria we use to select a system).

124. WHITTAKER ET AL., *supra* note 18, at 28; Bornstein, *supra* note 62, at 570 ("[A]ny improvement to traditional decision-making that relies on data will depend on what data is being used and how."); Chander, *supra* note 39, at 1036 (arguing that the main problem with algorithmic discrimination is that the algorithms learn from data that already shows discriminatory effects).

unlike human decision-makers, algorithms can separate information collection from processing. Instead of blocking information from the algorithm, we can allow it to collect information and then decide whether to use it—for example by making the algorithm compare the potential decision with information and the counterfactual decision without information.[125]

In other words, with an automated decision-making process, it is possible to collect all features and instruct the model to ignore one of those features while making a prediction.[126] Despite the risk of amplifying bias when unchecked, if regulated appropriately algorithms can be productive for reducing discrimination. In some ways, we demand more from algorithmic decision-making than we do from the human type because algorithms present the possibility of doing so. They do this in two different ways.

The first is de-biasing. While it is true that human biases can be coded into the algorithm,[127] the process of coding them makes them more explicit. This means that unconscious biases might be detected by the same programmer who holds them, or by subsequent reviewers, and not all biases will necessarily be transferred to the code.[128]

The second is monitoring. The fact that algorithms are coded makes it easier to regulate algorithmic decision-makers than human ones, absent trade secrets. While faulty logic is only figuratively coded in human decision-makers, it is literally coded in algorithmic ones. While this is less of an advantage for algorithms that operate as a black box to decision-subjects, such as deep learning systems,[129] oftentimes algorithms present an opportunity for accountability.[130]

Because algorithmic discrimination is a problem of managing humans intermediated by an algorithm,[131] it is not surprising that algorithmic

125. Matt Kusner et al., *Counterfactual Fairness*, 30 ADVANCES IN NEURAL INFO. PROCESSING SYS. 4066, 4075 (2017), http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf. One way to implement this could be to train a model with the information of the protected category in the training data, then train a second model without such information in the training data, and compare the results of both models with each other through a difference in difference regression, and each result with different definitions of algorithmic fairness. This would allow, for example, for race to be blocked when training algorithm 1 and then use algorithm 2 to perform auditing steps to control for disparate impact in algorithm 1, while knowing who is white and who is black in the database; then, one can create different decision rules for black and white individuals that could be more accurate than a one size fits all rule. However, this would run into the disparate-impact disparate-treatment tension described in *infra* Subpart IV.A.

126. While possible, this may be ineffective unless the feature involved is independent, as discussed in *infra* Part III.

127. *See generally* Winner, *supra* note 64 (describing how technologies can be designed with patterns and biases within them).

128. *See* Bornstein, *supra* note 62, at 526; Barocas & Selbst, *supra* note 21, at 673–74, *Data-Driven Discrimination*, *supra* note 21, at 869–71; *see also* Stephanie Bornstein, *Reckless Discrimination*, 105 CALIF. L. REV. 1055, 1058–59 (2017) (arguing that not using technologies, such as algorithms, to check subjective biases in hiring decisions should constitute discrimination under Title VII's disparate impact doctrine).

129. *See infra* Subpart IV.D.

130. *See* Ignacio Cofone & Katherine Strandburg, *Strategic Games and Algorithmic Secrecy*, 64 MCGILL L.J. (forthcoming 2019) (evaluating when algorithms ought to be kept secret due to gaming and when gaming is not an obstacle for disclosure).

131. *See supra* Subpart I.B.

discrimination and human discrimination have a similar dynamic regarding information flows. Whether a person making the ultimately discriminatory decision did so intermediated by an algorithm is only relevant for determining *how*, and not *whether*, to regulate information to prevent the discriminatory act.

The three conditions for the effectiveness of antidiscriminatory information rules previously described—skewed samples, processing problems (bias), and intent—[132] are also applicable to algorithms. Skewed samples to make inferences produce biased outcomes in algorithms as they do in humans.[133] Similarly, a biased process (with prior beliefs that filter into the code) and discriminatory intent are possible from the people that are involved in designing and applying the algorithm.[134] They can also (unknowingly to their designers) produce discriminatory outcomes when working with a database containing embedded social biases.[135]In other words, the algorithmic biases that can lead to discrimination (bias in the data that are used as an input and bias in the data-processing mechanism) trace back to the effectiveness conditions for antidiscriminatory information rules.[136]

The privacy solution to algorithmic discrimination, in this sense, is also to apply an information policy to algorithms, rather than only to human discrimination. However, there is an obstacle to regulating data to prevent bias and discrimination. As the next Subpart explores, blocking information is likely to make algorithmic discrimination only worse. Subpart D will then explore different formulations of the idea of never blocking information, and what this idea means for the question of how to prevent discrimination

D. THE KEY CHALLENGE FOR PRIVACY RULES: THE ENDLESS LINE OF PROXIES OBJECTION

When evaluating antidiscriminatory information rules, it is crucial to understand how systems identify proxies for information points that are proscribed by the law. Shifting discrimination from an information point to proxies is only possible when proxies for the blocked piece of information are available—especially when they are accurate and easily observable. The use of

132. *See* Subpart II.A. *See generally* Cofone, *supra* note 11.

133. *See* Subparts II.B, II.C.

134. *See supra* Subpart I.B; *see also Data-Driven Discrimination*, *supra* note 21, at 884 (discussing how employers can engage in intentional discrimination by "rel[ying] on an algorithm . . . because it *knows* the model produces a discriminatory result and *intends* that result to occur. . . . masquerade[ing] behind the neutral façade of data analysis . . . . [where] the pretext—the 'legitimate business reason' given for the decision—is the output of a computer model."); Barocas & Selbst, *supra* note 21, at 692–93 (noting that employers can mask intentional discrimination by using an apparently neutral algorithm).

135. Algorithms can also be used by people with a discriminatory intent who try to do proxy laundering: choosing seemingly inoffensive proxies or facially neutral rules with the coveted intention of discriminating against a protected category. But this is not an algorithmic problem. Instead, it is a problem that antidiscrimination law has been addressing for as long as disparate impact discrimination has existed. The main challenge (besides additional probatory difficulties) that algorithms pose for antidiscrimination law is not intentional discrimination, but unintentional discrimination.

136. *See generally* Cofone, *supra* note 11.

those proxies by the system can have an equally negative or even more devastating outcome than would have resulted from using the blocked information in the original discriminatory decision. Therefore, it is crucial for any effective antidiscriminatory information rule to identify and block those information flows as well. Checking for the existence of these proxies is necessary for predicting the effectiveness of any antidiscriminatory information rules and of crucial importance for designing better ones. However, it is important to keep in mind that this is rarely possible with algorithms. The proxies that algorithmic processes might identify, or even the fact of whether an algorithm will identify a proxy at all, is difficult—and sometimes impossible—to predict.

To apply an information policy to algorithms, a crucial task is to identify which pieces of information are proxies for others.[137] Blocking proxies for protected categories may be key for avoiding discriminatory outcomes.[138] However, two central problems have been identified for doing that. The first problem is that we may not know which those proxies are and, if we did, it may be impossible to block all proxies. The second problem is that, even if it is possible to block proxies, it may be undesirable as those proxies could also contain valuable information.

The first problem is perhaps the most commonly heard objection in the industry. An algorithmic process could produce a disparate impact on a protected category due to correlations between pieces of information that were initially hard to predict.[139] The reason for this is that a large number of features could correlate with the protected category only slightly, none of them altering the prediction significantly but all of them in aggregation doing so,[140] so the removal of any one of them would make no significant difference. The Amazon hiring algorithm serves as an example of this as well. After realizing that the algorithm discriminated based on gender, Amazon modified it to ignore words that denoted gender. However, the algorithm continued to "guess" individuals' gender by using other words in the resumes that correlated with gender.[141] Moreover,

137. *See infra* Subpart III.A.

138. Barocas & Selbst, *supra* note 21, at 691; Bodie et al., *supra* note 16, at 1022–23.

139. Bodie et al., *supra* note 16, at 1023 (arguing that the use of algorithms risks other pieces of data being proxies for prohibited categories); Benjamin Alarie et al., *How Artificial Intelligence Will Affect the Practice of Law*, 68 U. TORONTO L.J. 106, 116–17 (Supp. 2018) ("Machine learning's agnostic approach—choosing an algorithm that maximizes predictive accuracy independent of underlying theory—enables it to leverage connections between and among references, even those that are implied rather than expressed."); Hu, *supra* note 16, at 641, 695 (explaining that algorithms can use data—like risk factors—that are not protected categories, but serve as proxies for protected categories).

140. *See, e.g.*, Michal Kosinski et al., *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROC. OF THE NAT'L ACAD. OF SCI. OF THE U.S. 5802, 5805 (2013) ("[A] wide variety of people's personal attributes, ranging from sexual orientation to intelligence, can be automatically and accurately inferred using their Facebook Likes. Similarity between Facebook Likes and other widespread kinds of digital records, such as browsing histories, search queries, or purchase histories suggests that the potential to reveal users' attributes is unlikely to be limited to Likes. Moreover, the wide variety of attributes predicted in this study indicates that, given appropriate training data, it may be possible to reveal other attributes as well.").

141. REUTERS, *supra* note 42.

proxies can change their meaning over time and the proxies involved could be emergent: they may not be proxies before, but appear later in the process (a piece of information that was not a proxy for race in the past could become one in the future).

The corollary problem is that, if one wanted to block all proxies for protected categories, one would never cease to find more information points that, to some degree, are predictive of each other and would need to be blocked.[142] In that endeavor, one might have to block information *ad infinitum*. Moreover, in large databases, many of these proxies could be redundant with each other.[143] The potential that machine learning has for identifying new proxies is not a quirk, but the main interest in A.I. In other words, one could potentially design a system that blocks all proxies and develop neutral hiring practices if it had no information, but that would defeat the point of having such system.[144] What would be left is randomness.

The second problem is that, when one removes any information, one also takes away relevant information for decision-making.[145] A feature that is a proxy for a protected category could also be a proxy for useful and legitimate information, implying that blocking information presents a tradeoff.[146] For example, education may be predictive of race in some social contexts, but it is also predictive of job performance.

More importantly, to achieve standards of algorithmic fairness, one must be aware of the way in which the model's variables have differing predictive power among different protected groups.[147] In order to assess whether the algorithm has resulted in disparate impact discrimination, it is essential to know the value of the sensitive variable. For example, if the algorithm is "race blind"

142. Toshihiro Kamishima et al., *Fairness-Aware Learning Through Regularization Approach*, *in* 2011 IEEE 11TH INT'L CONF. ON DATA MINING WORKSHOPS 643, 643 (2011) ("[T]he simple elimination of sensitive features from calculations is insufficient for avoiding inappropriate determination processes, due to the indirect influence of sensitive information.").

143. Solon Barocas et al., Fairness and Machine Learning: Limitations and Opportunities 41 (Sept. 4, 2018) (unpublished textbook), http://fairmlbook.org ("[S]everal features that are slightly predictive of the sensitive attribute can be used to build high accuracy classifiers for that attribute. In large feature spaces sensitive attributes are generally *redundant* given the other features.").

144. Barocas & Selbst, *supra* note 21, at 675 ("Even in situations where data miners are extremely careful, they can still effect discriminatory results with models that, quite unintentionally, pick out proxy variables for protected classes."). In addition, the authors note:

> Cases of decision making that do not artificially introduce discriminatory effects into the data mining process may nevertheless result in systematically less favorable determinations for members of protected classes. This is possible when the criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership.

Id. at 691. *See generally* Jane Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1 (2018) (explaining how algorithmic decision-making works with proxies in a dynamic setting).; James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV ONLINE 164 (2016).

145. *Data-Driven Discrimination*, *supra* note 21, at 897-901. *See generally* Cynthia Dwork et al., *Fairness Through Awareness*, *in* PROC. OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONF. 214 (2012), http://doi.acm.org/10.1145/2090236.2090255; Grimmelmann & Westreich, *supra* note 144.

146. Grimmelmann & Westreich, *supra* note 144, at 171.

147. Dwork et al., *supra* note 145, at 18.

and the category of race is removed from the input data, then it will be impossible to determine whether the output is discriminatory on the basis of race. Therefore, blocking information may not only reduce accuracy but could also be self-defeating by reducing the ability to detect bias.[148]

This objection is also sometimes raised by industry members. The Google whitepaper on *Perspectives on Issues in AI Governance*, for example, states that:

> [I]nferring race can be essential to check that systems aren't racially biased, but some existing laws around discrimination and privacy can make this difficult. Similarly, while it might seem sensible to bar the inference of a person's gender to guard against unfair treatment, in practice doing so could inadvertently have the opposite effect, by making it harder to deliver reliable 'mathematically fair' gender-neutral outputs. We urge policymakers and experts to work together to identify where this kind of inadvertent counter-intuitive harm arises."[149]

Consequently, the first step to remove discrimination, according to this line of reasoning, is to increase the algorithm's accuracy.[150] The reason is simple: accuracy decreases error rates that are often discriminatory towards disadvantaged groups, as Strahilevitz argues for statistical discrimination. Collecting more data, and especially data about protected categories, is therefore a useful first step to reduce discrimination.[151]

In sum, because of machine learning algorithms' computational capacity, simply blocking information like the law does with human subjects may more often than not be impossible, ineffective, or undesirable. When aiming to prevent race discrimination, for example, one may find that a host of attributes in the United States correlate in some way with race. Some of those may be unexpected, and some of those may be legitimate considerations for the decision. This idea has dominated much of the discourse around algorithmic discrimination: blocking information about protected categories will be ineffective, so there is nothing that one can do ex-ante to prevent algorithmic discrimination. The next Part qualifies this position.

148. *Data-Driven Discrimination*, *supra* note 21, at 917–18.

149. GOOGLE, PERSPECTIVES ON ISSUES IN AI GOVERNANCE 15 (2018).

150. Accuracy can be defined as the proportion of examples of which the model produces a true positive or true negative result. GOODFELLOW ET AL., *supra* note 31, at 101–02; *see also* Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE ACCESS 52138, 52141, tbl. 1 (2018) (defining accuracy as the "performance metric that refers to the number of correct predictions made by the model" over all predictions made).

151. Kroll et al., *supra* note 15, at 685; *see, e.g.*, Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually Not that Clear,* WASH. POST, (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

III. FOCUS ON DATA REGULATION, NOT ALGORITHMIC REGULATION

Laws that govern A.I. can play a significant role in how A.I. learns and acts in the world.[152] This Article showed so far that the law often regulates information to prevent discrimination but does so with varied effectiveness because of the sources of human bias.[153] Algorithmic bias, on the other hand, is based directly on the information that the algorithms are fed. Therefore, their bias can be addressed by changing their information input. But blocking information on protected categories may be ineffective because an infinite number of data points will be proxies for them.

Since regulating the availability and types of information points available is essential to regulating algorithmic processes, this Part shows that not all proxies are harmful and sometimes a change from an information point to its proxy is desirable. It then shows when blocking information does not fall under the objection about the endless line of proxies. Finally, and most importantly, it shows that changing data input does not necessarily mean blocking information about protected categories. Rather than blocking information, we can either encode or shape it. The first two Subparts qualify the objection that it is impossible to block all proxies for protected categories and show that the objection's range of applicability is narrower than often assumed. The third and fourth Subparts show how to overcome the problem when the objection applies.

A. NOT ALL PROXIES ARE BAD PROXIES

When an information point is blocked, decision-makers who try to gauge it, algorithmic or human, are forced into using different information that may serve as a proxy. The three categories of proxies that an algorithm can be forced into are transfer proxies, reducing proxies, and expanding proxies.[154] The difference between them depends on the relationship between the baseline population (referred to as the 'information-point' group) and the targeted group when the proxy is used (the 'proxy group').

Transfer proxies lead algorithms to, instead of targeting the protected 'information-point' group, target a different group that has some overlap with the 'information-point' group (see Figure 1). As a result, the subset of the 'information-point' group that does not overlap with the new proxy group will be protected from discrimination, and the subset that overlaps between both groups will see no change in their situation. However, the subset of the 'proxy group' that does not overlap with the 'information-point' group will see their situation worsened: they will face discrimination while they did not before.[155]

152. *See generally* Levendowski, *supra* note 16.
153. Cofone, *supra* note 11, at 150.
154. Cofone, *supra* note 11, at 154–58.
155. *Id*. at 155.

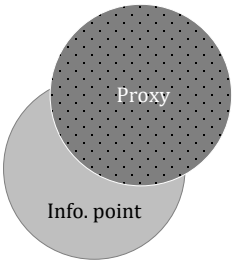

Figure 1: Illustrates transfer proxies

Reducing proxies narrow down discrimination to a subset of the ‘ìnformation-point’ group (see Figure 2). Unlike transfer proxies, they do not target new individuals. Instead, they benefit some members of the ‘information-point’ group (those that are not part of the ‘proxy group’), and leave the members of the ‘proxy group’ in the same situation as before the proxy was used.[156]

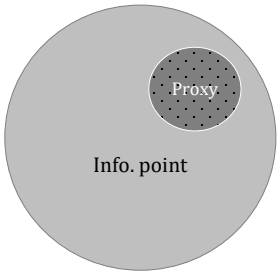

Figure 2: Illustrates reducing proxies

Expanding proxies have the opposite effect of reducing proxies (see Figure 3). The ‘information-point group,’ for expanding proxies, is a subset of the ‘proxy group’. As a result, the target of discrimination is difussed among members of the ‘proxy group’ in a probabilistic fashion, instead of being focused on members of the ‘information-point’ group.[157]

156. Imagine an industry with an aversion towards hiring Latinos. Imagine that, although being Latino is unobservable, Spanish-sounding first or last names serve as proxies. To mitigate discrimination, policy-makers could enforce a system that requires resumes to only contain initials for first names. As a result, decision-makers would only be able to use last names as proxies for being Latino. This would reduce discrimination from the larger ‘Spanish-sounding first name or last name’ information group to the smaller ‘Spanish-sounding last name’ proxy group. Indiivduals with a Spanish-sounding first name but without a Spanish-sounding last name would be protected from discrimination. This would not solve the problem, as individuals with Spanish-sounding last names will still be targets of discrimination, but it would create an improvement over the prior situation. *Id*.

157. For example, imagine an employer who wants to avoid hiring someone who might take a maternity leave in the near future. Most jurisdictions prohibit asking this question. Many scholars argue that this prohibition may lead employers to disadvantage all women. However, the employer will not disadvantage this larger group of candidates as much as it would have disadvantaged candidates whom it asks about maternity

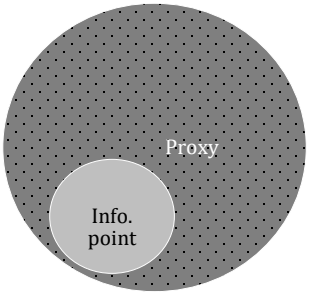

Figure 3: Illustrates expanding proxies

When proxies for protected categories blocked from decision-makers are available, blocking them as well may seem like the most effective method to prevent discrimination. There are few situations in which the problematic data will be what computer scientists call an independent feature of the model: an information point that is not significantly correlated with other features or information points. Put differently, sometimes the information on an individual's membership of a protected category, or one of its proxies, will not be correlated with other pieces of information that the algorithm uses. This situation will be infrequent. If the protected category is an independent feature, then blocking the protected category is possible without falling into the problem of the endless line of proxies.[158] Traditional statistical methods allow us to determine when this is the case.[159]

But more importantly, some proxies, while impossible or unfeasible to block, may be inoffensive or may improve the situation compared to one in which an algorithm uses information about a protected category. Even if there is an endless line of proxies for the protected category, some of those may diffuse or reduce the discrimination in a way that makes it desirable to shift the algorithm into those proxies.

Shifting discrimination from one population to another through transfer proxies is generally undesirable as it harms a new group of individuals. However, antidiscriminatory information rules can be designed to reduce the size of the targeted group through reducing proxies. Alternatively, expanding proxies can be used to diffuse discrimination among a larger group in a probabilistic fashion, and avoid targeting any one group directly.[160]

Several considerations will determine whether the use of reducing or expanding proxies is desirable. It would be undesirable to use expanding or reducing proxies if the proxy group is one that has been historically protected by legislatures or is particularly vulnerable—for example because it is

leave and obtains a positive answer; without asking being allowed, any of these candidates will only have a probability of going on maternity leave. *See id.* at 156–57.

158. *See supra* Subpart II.D.

159. For example, one could run a Chi-square test with attention to how high the power is.

160. Cofone, *supra* note 11, at 157.

intersectional. For expanding proxies, because the process is probabilistic, the relative sizes of the proxy groups matter; these proxies will be useful when the information is diffused among a larger group (ideally the entire population).[161] It will also be desirable to make an algorithm shift to these proxies when traditional antidiscrimination law is more effective at addressing the remaining discrimination against the proxy group—for example because it is larger.[162]

In other words, for antidiscriminatory information rules to be effective, they need not block all proxies.[163] They must instead identify proxies that reduce or expand the group of people that are discriminated against and gauge those proxies' usefulness in helping protected categories.[164]

B. WHEN TO BLOCK INFORMATION DESPITE PROXIES

The last Subpart showed that it may sometimes be desirable to make decision-makers shift from an information point to some types of proxies. However, even when there are undesirable proxies, it may be desirable to block information on protected categories.

The salient characteristic of machine learning algorithms for antidiscriminatory information rules is that, for these algorithms, there is evidence of the ineffectiveness of blindness as a fairness principle.[165] This problem exists because machine learning algorithms are more emergent than knowledge-based systems: the outcome is more difficult to predict for the humans that make and apply them.[166]

For human decision-making, the law and social norms use blindness frequently. For example, in a double-blind peer review, reviewers cannot see the personal characteristics of the author whose paper they are reviewing in the hope

161. In the maternity leave example, it is unlikely that an employer will stop hiring all women. However, it is possible that an employer will avoid hiring all members of a relatively smaller group, such as a religious minority group or an intersectional group. Consider a scenario where an employer wants to avoid hiring Muslim women who wear a veil. A policymaker who wants to devise information rules to protect them should evaluate how many Muslim women choose to wear a veil. If most do, then blocking the information might lead employers to use being Muslim as a proxy for wearing a veil, and shift discrimination to all Muslim women. Instead of protecting some women from discrimination, this shift would introduce discrimination towards every member of a larger group. Contrary to the policy's aims, it would worsen the situation. The question of whether to make the probabilistic shift in these two examples may appear similar, but the groups' relative sizes lead to opposite policy conclusions. *See Id.*

162. *Id*. at 157–58.

163. *See* Cofone, *supra* note 11.

164. *See id.*

165. Dwork et al., *supra* note 145, at 7. Blindness is the idea of not looking at the protected attribute when making a decision.

166. Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 538–45 (2015) (explaining emergence as one of the disruptive characteristics of robotics for the law); Jack M. Balkin, *The Path of Robotics Law*, 6 CALIF. L. REV. CIRCUIT 45, 45–51 (2015) ("I do not distinguish sharply between robots and artificial intelligence (AI) agents . . . . We may be misled if we insist on too sharp a distinction between robotics and AI systems . . . ."); Cofone, *supra* note 14, at 185 ("[E]mergence will determine the extent to which an A.I. agent's behavior is foreseeable to people who have free will and who live under the rule of law—which is fundamental to determining liability under tort law.").

that such rule will lead them to be more objective. Similarly, at McGill University we blind all exams with an anonymous exam code to help professors be more objective while grading. This idea can be easily extended to knowledge-based systems. The effectiveness of these measures can be persuasively contested, as can their applicability to other contexts, but for human decision-makers and knowledge-based systems, it at least makes some sense.

For machine learning algorithms, it seemingly does not. In machine learning, when there is a fair sample and no processing problems, algorithmic bias is often a side-effect of maximizing accuracy with regards to the chosen variable.[167] When bias is a side-effect of accuracy, unlike prejudice, it functions like statistical discrimination. Recall the example above of people with green hair and purple hair.[168] If there is differential observability, where people with purple hair have less traditional educational credentials than people with green hair and, therefore, less observable proxies for skill, maximizing accuracy will lead to disproportionately hiring people with green hair.

Moreover, machine learning is much better at detecting proxies from the dataset than is any human decision-maker. We may block hair color from the database, but a machine learning algorithm may induce it from a number of data points that correlate to it.[169]

The natural exception to this rule, then, is when the information does not improve accuracy. A situation where one should block information from a machine learning model, for example, is label leakage, where some of the input data filter into the label in a way that they reduce the model's accuracy.[170] Take a real-world example from a cancer prediction machine learning model. The model was trained to predict the probability of a patient having cancer based on his or her medical records.[171] The algorithm picked up features from these records such as age, test results, and gender. Although the model had satisfactory predictions on the test data,[172] it performed poorly when applied to new

167. *See supra* Subpart II.C. *See generally* Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, 30 ADVANCES NEURAL PROCESSING SYS. 1 (2016) http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf (quantifying this idea with simulated data baseed on FICO scores).

168. *See supra* Subpart II.D.

169. Barocas et al., *supra* note 143, at 41 ("Some have hoped that removing or ignoring sensitive attributes would somehow ensure the impartiality of the resulting classifier. Unfortunately, this practice is usually somewhere on the spectrum between ineffective and harmful. In a typical data set, we have many features that are slightly correlated with the sensitive attribute. Visiting the website pinterest.com, for example, has a small statistical correlation with being female.").

170. Marzyeh Ghassemi et al., *Opportunities in Machine Learning for Healthcare* 4 (2018) (unpublished manuscript), https://arxiv.org/pdf/1806.00388.pdf; Truyen Tran et al., *Preterm Birth Prediction: Deriving Stable and Interpretable Rules from High Dimensional Data*, *in* PROC. OF MACHINE LEARNING FOR HEALTHCARE 2016 at 3 (2016) http://proceedings.mlr.press/v56/Tran16.pdf.

171. Shachar Kaufman et al., *Leakage in Data Mining: Formulation, Detection, and Avoidance*, *in* PROC. OF THE 17TH ACM SIGKDD INT'LCONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 556, 557 (2011), https://www.cs.umb.edu/~ding/history/470_670_fall_2011/papers/cs670_Tran_PreferredPaper_LeakingInDataMining.pdf.

172. Having held out data to keep training data separate from testing data, as is best practice.

patients.[173] It turns out that one of the features that the model learned from patients' medical records was the hospital where they were being treated.[174] Many hospitals specialize in caring for cancer patients, so these hospitals were highly predictive of whether the patient had cancer: being treated in a hospital that specializes in cancer treatment highly correlates with having cancer.

This effect was increased for institutions with the word "cancer" in it, but the first effect prevailed even when anonymizing hospital names.[175] The new patients that this model was designed to be applied to, however, lacked this information. The aim of the model was to use it to determine the probability of a patient having cancer for it to help in the allocation of patients among hospitals. Showing the model the hospital names was cluing the model into a doctor's diagnosis that the new patients, who the model was supposed to be applied to, lacked.[176] The model would have performed better in its predictions without such information. Label leakage in this case led to problems identifying cancer patients, but it could also lead, for example, to problems identifying productive employees if a model took labeled features in a similar way.

In sum, blocking information on protected categories when there are proxies available is rarely a first best solution but, in some situations, such as label leakage, blocking information may nonetheless be desirable.

C. LEARNING FROM UTOPIA: SHAPING THE DATA

As mentioned above, there are few situations in which the problematic data may be an independent feature of the model. But we should assume the most difficult, and most frequent, case: that the protected category is deeply linked to a long line of proxies. There is an alternative between unhelpfully blocking sensitive attributes and passively allowing for the discrimination they may produce: not blocking the data, but altering them.

I suggest two ways of doing this. The first way, explored in this Subpart, is shaping the data. In some way, this is to "lie" to the algorithm, pretending that we live in the kind of society that we want to live in. The other, explored in the following Subpart, is encoding protected categories in the training data instead.

Imagine that Amazon had trained its machine learning algorithm with a gender balanced sample of its employees. Or imagine that the data fed into COMPAS had been racially balanced. They would have not produced their highly criticized results.

---

173. Kaufman et al., *supra* note 171, at 557–58; *see also* Claudia Perlich et al., *Breast Cancer Identification: KDD CUP Winner's Report*, 10 SIGKDD EXPLORATIONS 2, 39–40 (2008) https://www.kdd.org/exploration_files/KDDCup08-P1.pdf (describing that the Patient ID input variable sometimes contained information that resulted in label leakage, such as the name of the institution, or the type of medical advice).

174. *Id.* at 561–62.

175. *Id.*

176. *Id.*

Note that COMPAS was not biased as a measure of re-arrest; it was biased as a measure of re-offense.[177] COMPAS used its prediction of re-arrest as a proxy for re-offence. In doing so, it picked up the social biases that distort the relationship between offending and being arrested. Similarly, the Amazon employment algorithm was not biased in measuring Amazon's historical hiring practices. It was biased as a measure for Amazon's best job candidates because social biases interfered in the relationship between Amazon's historical hires and Amazon's current best job candidates. In other words, the problem is the human decision of using re-arrest as a proxy for re-offending and past hires as a proxy for best hires. This is what generates a statistical bias, or a difference between the estimator's expected value and its true value, similar to the simplified problem above between people with green hair and people with purple hair.[178] Because the choice of proxy generated the statistical bias, it is reasonable to require that statistical measures are taken to correct for such bias if the decision falls under disparate impact discrimination.

Because disparate impact discrimination is a problem of adversely affecting protected populations without a classification bias (which is a bias in the process), it is, in a way, a data input problem. Unlike humans, algorithmic bias is based directly on the information that the algorithms are fed: they do not keep irrational impermeable biases, but their biases are heuristics.[179] To address these biases, companies and government agencies could be required to modify the training data.[180] In other words, to conform with antidiscrimination law, they could be required to consider that, even though groups are not equal in the real world, they must be treated equally for the purposes of the decision-making process, and could be asked to train their models to conform to such belief. This can be achieved through pre-processing, or sanitizing, the data with which a machine learning model is trained.[181]

---

177. Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, *in* PROC. OF THE 23RD ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 797, 803, fig. 2 (2017) (noting that calibration of COMPAS was satisfactory in the sense that the predictive accuracy of re-arrest was the same for both groups).

178. *See supra* Subpart I.E.

179. Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 191 (2017) (explaining that bias in machine learning is caused by social processes that are reflected in the data).

180. *See, e.g.*, Flavio P. Calmon et al., *Optimized Pre-Processing for Discrimination Prevention*, *in* 30 ADVANCES NEURAL INFO. PROCESSING SYS. 3993 (2017), https://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf (proposing a probabilistic framework for discrimination-preventing preprocessing that results in fairer classifications with slightly reduced accuracy); Indrė Žliobaitė, *Fairness-Aware Machine Learning: A Perspective* (Aug. 3, 2017) (unpublished manuscript), https://arxiv.org/pdf/1708.00754.pdf (arguing that preventing discrimination requires an understanding of how it appears in machine learning and expressing the concern that, without such understanding, too overly strict regulations may occur).

181. Pre-processing approaches, which modify the source data, exist in opposition to in-processing approaches, which modify the algorithm to add antidiscrimination constrains, and post-processing approaches, which fix the resulting model. *See* Brian d'Alessandro et al., *Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification*, 5 BIG DATA 120, 129–30 (2017) (explaining such distinction); Sara Hajian and Josep Domingo-Ferrer, *A Methodology for Direct and Indirect Discrimination Prevention in*

While the purpose of this Article is to propose to regulate information to prevent algorithmic discrimination, and not to develop technical methods to achieve this, it is important to note that these methods exist. Building on this knowledge, we can find an alternative in between unhelpfully blocking information about protected categories and passively allowing for disparate impact discrimination.

There are different technical ways of addressing the distribution of data to train machine learning algorithms. One such method is developing a separate, tunable, de-biasing algorithm that adjusts the sampling probabilities of each data point during the training stage: it learns the structure of the data and changes the weights of different data points during training.[182] Because probabilities are adjusted during training, such method can also aid with unknown biases in the training data without the need to label protected categories in such data.[183] This method has been tested, for example, on facial detection algorithms achieving both an increase in accuracy and decrease in race and gender biases.[184] Re-sampling and re-weighting are similar. Re-sampling compares the expected size of a group to its actual size and samples accordingly, possibly duplicating data points, to achieve a fair distribution.[185] Reweighing, similarly, changes each individuals' weight to neutralize inequalities embedded in the historical data.[186]

Data Mining, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENG'G 25, no. 7, 1445 (2013) (proposing data mining systems that are discrimination-conscious by-design). Some in-processing techniques are naïve Bayes models, logistic regression, hinge loss, support vector machines, adaptive boosting, decision trees, and classification. Some post-processing techniques are classification and rule & pattern mining. *See* Toon Calders & Sicco Verwer, *Three Naive Bayes Approaches for Discrimination-Free Classification*, 21 DATA MINING KNOWLEDGE DISCOVERY 277, 288–90 (2010) (utilizing naive Bayes models); Muhammad Bilal Zafar et al., *Fairness Constraints: Mechanisms for Fair Classification* (Mar. 23, 2017) (unpublished manuscript), http://arxiv.org/abs/1507.05259 (utilizing logistic regression, hinge loss, support vector machines); Benjamin Fish et al., *A Confidence-Based Approach for Balancing Fairness and Accuracy*, *in* PROC. OF THE 2016 SIAM INT'LCONF. ON DATA MINING 144 (Sanjay Chawla Venkatasubramanian & Wagner Meira eds., 2016), https://epubs.siam.org/doi/abs/10.1137/1.9781611974348.17 (utilizing logistic regression, support vector machines, adaptive boosting); Faisal Kamiran et al., *Discrimination Aware Decision Tree Learning*, *in* 2010 IEEE INT'LCONF. ON DATA MINING 869 (2010) (utilizing decision trees); Richard Zemel et al., *Learning Fair Representations*, 28 PROC. MACHINE LEARNING RES. 325 (2013) (utilizing in-processing classification). *See* Dwork et al., *supra* note 145 (applying post-processing classification); Faisal Kamiran et al., *Decision Theory for Discrimination-Aware Classification*, *in* 2012 IEEE 12TH INT'LCONF. ON DATA MINING 924 (2012) (applying post-processing classification); Dino Pedreschi et al., *Measuring Discrimination in Socially-Sensitive Decision Records*, *in* PROC. OF THE 2009 SIAM INT'LCONF. ON DATA MINING 581–92 (Chid Apte et al. eds., 2009) (applying rule and pattern mining); *see also* Lehr & Ohm, *supra* note 17, at 683 (describing a data cleaning stage in their eight-stage model of machine learning, although focusing their description on cleaning for accuracy and missing values).

182. Alexander Amini et al., *Uncovering and Mitigating Algorithmic Bias Through Learned Latent Structure*, *in* AIES CONF. (2019), http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_220.pdf.

183. *Id*.

184. *Id*.

185. *See* Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification Without Discrimination*, 33 IN KNOWLEDGE & INFO. SYS. 1 (2012) (developing three pre-processing techniques: suppressing the sensitive attribute, changing class labels, and reweighing or resampling data).

186. *Id*.

By pre-processing data, it is possible to modify the historical data that contain embedded inequality and translates it to models trained with it that would amplify it, to strip such data from discrimination. Pre-processing, in other words, modifies the training data to strip them from embedded bias that leads to disparate impact discrimination, to then train the model with fair data. This can be done either by modifying the input data generally,[187] or by modifying the target variable (race, gender, etc.).[188]

We saw above how blocking protected categories is ineffective to prevent algorithmic discrimination.[189] However, that does not mean that having more data is always better. If we are concerned with algorithmic disparate impact, and we believe that antidiscrimination law may not be equipped to adequately address it ex-post, then we must regulate the data. To succeed in doing this, rather than more data or less data, we need *fair* data.

Fair data, counterintuitively, means data samples that are unrepresentative of the pool, because it looks like what we believe the pool *would* look like had it not embedded structural inequalities. Fair data, or more meaningful data means, in some way, a "biased" data sample that counterweights social biases.

D. ACTIONABLE PRIVACY: ENCODING THE DATA

For some types of algorithms, applying the abovementioned approach is difficult because the model keeps learning while it is applied. These models are usually called reinforced learning models. Reinforced learning adds another layer of complexity to the regulation of information because the model's learning power is not limited to training data.[190] It is difficult, therefore, to control what the algorithm learns if it is being applied "in the wild." A possible intermediate solution for reinforced learning algorithms, in between uselessly blocking labels for protected categories and unhelpfully allowing the algorithm to mine all information and reach disparate impact outcomes, is to encode the data. While this method is far from perfect, it allows for group fairness to be built into the system.

This idea builds on the consideration explored above that, without incorporating sensitive features, it is impossible to correct for the impact of those sensitive features in other parts of an individual's "data package."[191] The

187. *See* Michael Feldman et al., *Certifying and Removing Disparate Impact*, *in* PROC. 21ST ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 259 (2015); Zemel et al., *supra* note 181.

188. *See generally* Faisal Kamiran & Toon Calders, *Classifying Without Discriminating*, *in* PROC. 2ND INT'L CONF. ON COMPUTER, CONTROL AND COMMC'N (2009) (introducing a model trained with "biased" data using a ranking function, which works impartially with future data, reducing discrimination in future classifications); Koray Mancuhan & Chris Clifton, *Combating Discrimination Using Bayesian Networks*, 22 ARTIFICIAL INTELLIGENCE L. 211 (2014); Faisal Kamiran et al., *Quantifying Explainable Discrimination and Removing Illegal Discrimination in Automated Decision Making*, 35 KNOWLEDGE & INFO. SYS. 613 (2013).

189. *See supra* Subpart III.B.

190. *See generally* Shahin Jabbari et al., *Fairness in Reinforcement Learning*, 70 PROC. MACHINE LEARNING RES. 1617 (2017) (explaining the additional challenges of defining fairness in a reinforcement learning process).

191. *See supra* Subpart II.D.

algorithm, in other words, cannot be "race blind" and, at the same time, not engage in disparate impact discrimination.[192] Luckily, this is a problem that antidiscrimination law has been addressing for a long time. It is also a problem that the regulation of personal information for decision-makers has been addressing for a long time.[193]

As we do with human decision-makers, we can turn to different ways to obfuscate the data. For human decision-makers, these data are sometimes blocked: the law prevents decision-makers from accessing the information to prevent them from using it unfairly.[194] For algorithms, instead of being blocked, the data can be encoded.

The first way to do this is through what is referred to as actionable privacy. Machine learning research has shown that fair outcome-based models may be learned by incorporating, and at the same time encrypting, individuals' sensitive attributes.[195]

One way of implementing this is through multi-party computation. With such method, it is possible to encrypt attributes that denote membership of a protected category while keeping statistical data on protected categories.[196] The procedure is as follows. First, users encrypt their personal data. Then, they send it to the company.[197] After that, the company engages in multi-party computation alongside a regulatory body. In such way, neither the company nor regulator know all information at one time.[198] In other words, nobody can see all sensitive information.

Actionable privacy is different from differential privacy. A guarantee of differential privacy is the promise that, given the model, anonymization cannot be reverted; it is a model inversion guarantee.[199] For example, iOS uses differential privacy to report app usage to Cupertino: the system adds random noise so that Apple can collect the aggregated data for statistical purposes

192. Niki Kilbertus et al., *Blind Justice: Fairness with Encrypted Sensitive Attributes*, 80 PROC. MACHINE LEARNING RES. 2630 (2018).

193. *See supra* Subpart III.A. *See generally* Cofone, *supra* note 11.

194. Cofone, *supra* note 11, at 11.

195. Kilbertus et al., *supra* note 192; *see also* Michael Veale & Reuben Binns, *Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data*, BIG DATA & SOC'Y, Nov. 20, 2017, at 5 (proposing to use third parties that hold individuals' sensitive data); James E. Johndrow & Kristian Lum, *An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction* 11–12 (Mar. 16, 2017) (unpublished manuscript), https://arxiv.org/abs/1703.04957v1 (proposing a method to eliminate bias from predictive models by removing all information regarding protected variables from the training data and applying such method to COMPAS dataset).

196. *See generally* Kilbertus et al., *supra* note 192.

197. *Id.* at 2632.

198. *Id.* at 2632-34.

199. CYNTHIA DWORK & AARON ROTH, THE ALGORITHMIC FOUNDATIONS OF DIFFERENTIAL PRIVACY 1 (2014); Kobbi Nissim et al., *Bridging the Gap Between Computer Science and Legal Approaches to Privacy*, 31 HARV. J.L. & TECH. 687, 713–33 (2018), https://dash.harvard.edu/handle/1/37355739 (applying Dwork's differential privacy model to the Family Educational Rights and Privacy Act of 1974); Alexandra Wood et al., *Differential Privacy: A Primer for a Non-Technical Audience*, 21 VAND. J. ENT. & TECH. L. 209, 218–20 (2018).

without becoming aware of the identity of the app users.[200] With differential privacy, but not with actionable privacy, information about protected categories that could be helpful to detect discrimination is lost. In other words, differential privacy is the algorithmic equivalent of blocking information to algorithms like we do with humans, while actionable privacy is a method for encoding it instead.

Another way to achieve some degree of obfuscation is through fair representation. Computer science research has developed means of "finding an intermediate representation of the data that best encodes the data (i.e. preserving as much information about the individual's attributes as possible) while simultaneously obfuscat[ing] aspects of it, removing any information about membership with respect to the protected subgroup."[201] Put differently, it develops a representation of the data that preserves the information that the algorithm needs while encoding sensitive attributes. That allows the algorithm to capture useful information on group identity while at the same time blinding the process as to whether *each* individual is a member of the protected category through encoded representation.[202] Each individual is mapped in a probability distribution of a new dataset that ignores any information about whether the specific individual belongs to the protected category, while keeping group information about the protected category and satisfying statistical parity.[203] Obfuscation can also be achieved by adding noise to the data sample so that it is more difficult to predict protected class membership for each individual from the values of the different input variables.[204]

A biased machine learning algorithm (independently of whether it has reinforced learning) is so because it picked up biases from the training data.[205] Because algorithms are trained by randomly splitting available data into training data and testing data, algorithms may appear unbiased when evaluated in the testing data simply because the testing data and the training data contain the same biases.[206] For an antidiscriminatory information policy to be effective, we must collect and control information. We need to collect information on race in order to see impact on race, but we must also prevent information on race from producing discrimination based on race.

200. APPLE, DIFFERENTIAL PRIVACY TECHNICAL OVERVIEW, https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf (last visited July 27, 2019).

201. Zemel et al., *supra* note 181, at 326.

202. *Id.* (adding that this can be applied to any black box algorithm by applying the encoded classifier to the sanitized dataset).

203. *Id.* at 325.

204. Feldman et al., *supra* note 187, at 9–11.

205. *See supra* text accompanying notes 130–131.

206. Antonio Torralba & Alexei A. Efros, *Unbiased Look at Dataset Bias*, *in* 2011 IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION 1521, 1524–25 (2012) (proposing the idea of cross dataset generalization: a way of seeing if a dataset is biased is to train a model with it and then run the model in a wider and more diverse dataset to see if its accuracy drops).

IV. DOCTRINAL AND POLICY CONSEQUENCES

The previous Part demonstrated that we can deploy pre-processing techniques to shape or encode the training data, and it also suggested that we should do such. This Part explores the doctrinal and policy consequences of deploying pre-processing techniques to elaborate on why we should adopt such a proposal. This idea is built on the antisubordination theory developed by Fiss, Balkin, and Siegel. There are three ways in which an information approach to algorithmic discrimination is beneficial for antidiscrimination law: (i) correcting for the shortcomings of applying disparate impact protection without engaging disparate treatment, (ii) providing the possibility to include contexts in which disparate impact statues do not exist or are inapplicable to a certain minority group, and (iii) avoiding the problems that algorithmic opacity poses for traditional antidiscrimination law. At a policy level, by operating ex-ante, it also avoids the social harms created by discriminatory conduct, only partially solved by ex-post compensation.[207]

A. OVERCOMING THE DISPARATE-IMPACT-DISPARATE-TREATMENT TENSION

Disparate treatment forbids decision-makers from making distinctions based on protected categories, such as choosing not to hire women or black men. Disparate impact forbids them from making decisions that impact protected categories disproportionately, such as choosing only to hire people above a certain height (which adversely affects women) or people who can shave (which adversely affects black men). Unlike other jurisdictions where both disparate treatment and disparate impact are covered by antidiscrimination law,[208] in the U.S. disparate impact has a limited scope.[209] It is applied only when it is explicitly recognized by a statute,[210] such as Title VII or the Federal Housing Act.[211]

---

207. Austin, *supra* note 98, at 144.

208. Katerina Linos, *Path Dependence in Discrimination Law: Employment Cases in the United States and the European Union*, 35 YALE J. INT'L. L. 115, 131 (2010); Joseph A. Seiner, *Disentangling Disparate Impact and Disparate Treatment: Adapting the Canadian Approach*, 25 YALE L. & POL'Y. REV. 95, 117–20 (2006). These exist under the categories of direct and indirect discrimination (European Union) and direct and adverse effect discrimination (Canada).

209. Griggs v. Duke Power Co., 401 U.S. 424, 429–32 (1971) (holding that Title VII invalidates facially-neutral requirements with a disparate impact on a protected category even without discriminatory intent unless there is a proven relationship between requirements and job performance).

210. Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1, 4 (2006) (arguing we should move away from the statutory requirements and toward a structural approach to antidiscrimination law); Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 495 (2003); George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313, 2316 (2006); Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 732 (2006).

211. National Housing Act of 1934, 12 U.S.C. § 1701 (1934); *see also* Texas Dep't of Hous. and Cmty. Affairs v. Inclusive Communities Project, 135 S. Ct. 2507 (2014) (ruling that the Federal Housing Acts includes disparate impact claims).

Like traditional disparate impact doctrine, the algorithmic discrimination literature focuses on facially neutral practices that have disproportionately adverse effects on protected categories,[212] independently of whether there is discriminatory intent.[213] In algorithmic decision-making, classification schemes can be used to exacerbate inequality or disadvantage a protected category, but ignoring data about protected categories can also lead to disparate impact.[214]

This tension between disparate treatment and disparate impact is not exclusive to algorithms. In *Ricci v. DeStefano*, for example, courts faced the question of whether setting aside the results of a test to promote firefighters in the New Haven Fire Department to avoid a disparate impact outcome violated Title VII.[215] The Supreme Court held that it did.[216] So far, however, it has been possible for courts to find ad-hoc ways to argue around the tension. In *Ricci v. DeStefano*, the Court did so by arguing that the New Haven Fire Department lacked a "strong basis in evidence" to believe that it would have been otherwise held liable for disparate impact,[217] a workaround that led to criticism.[218] However, with algorithms this tension becomes more evident because ad-hoc workarounds turn difficult to implement due to scale.

This tension has led Barocas to develop the idea of unacknowledged affirmative action, arguing that it is difficult, if not impossible, to draw a clean line between concepts of algorithmic fairness and affirmative action when there is different prevalence among groups.[219] This difference can take place due to measurement bias. For example, using re-arrests as a proxy for recidivism when black individuals get arrested more than white individuals (maintaining amount of recidivism stable) will lead to more false positives for black individuals than for white individuals. However, it can also take place because of historical prejudice. For example, using advanced degrees as a proxy for intelligence when

212. *See generally* Kim & Scott, *supra* note 51 (discussing that while online targeting algorithms might use facially neutral variables, they can result in disparate impact as many of the facially neutral categories are proxies for categories like gender, age, or race).

213. *Id*. at 25 ("In many ways, discriminatory online targeting fits well with past disparate impact cases . . . courts today should find a disparate impact when employers target their recruitment ads using neutral attributes that disproportionately exclude users along the lines of race or other protected bases.").

214. Zachary C. Lipton et al., *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, 31 ADVANCES NEURAL INFO. PROCESSING SYS. 8136, 8138 (2018), http://papers.nips.cc/paper/8035-does-mitigating-mls-impact-disparity-require-treatment-disparity.pdf; Hu, *supra* note 16, at 645; *see also* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1198 (2017).

215. 557 U.S. 557 (2009). In the case, twenty firefighters sued after the city of New Haven invalidated a test because nineteen out of the twenty people chosen for a promotion based on the test were white.

216. *Id*.

217. *Id*. at 563.

218. *See, e.g.*, Mark S. Brodin, Ricci v. Destefano*: The New Haven Firefighters Case and the Triumph of White Privilege*, 20 S. CALIF. REV. L. & SOC. JUST. 161 (2011); Ann C. McGinley, Ricci v. DeStefano*: Diluting Disparate Impact and Redefining Disparate Treatment*, 12 NEV. L.J. 626 (2012); George Rutherglen, Ricci v. DeStefano*: Affirmative Action and the Lessons of Adversity*, 2009 SUP. CT REV. 83.

219. Solon Barocas, *What is the Problem to Which Fair Machine Learning is the Solution*?, AI NOW 2017 SYMPOSIUM (2017), https://ainowinstitute.org/symposia/videos/what-is-the-problem-to-which-fair-machine-learning-is-the-solution.html.

low-income individuals have fewer advanced degrees than high-income individuals (maintaining intelligence stable) will lead to more false negatives for low-income individuals than for high-income individuals.[220]

A fairness intervention can correct the algorithm's measurement bias. However, when correcting for a prevalence difference in an output, it is not always possible to know how much of it happens because of a measurement bias and how much happens because of historical prejudice. Therefore, Barocas argues, this would be to engage in affirmative action,[221] which is sometimes considered prohibited by the anticlassification principle.[222] Using information on membership of a protected category to treat its members differently could give rise to a disparate treatment challenge.[223]

Because the law forbids disparate treatment, it is often considered against the law to have race as a classifier in the decision; a decision-maker can collect information on race, but cannot decide differently based on race. Therefore, we cannot use a model that applies different cutoffs to different categories based on race to ensure an equal balance of false positives and false negatives. This prohibition to use race may harm, therefore, the very group that it is attempting to protect.[224]

This problem does not apply to the information policy proposed here because, while it addresses concerns of disparate impact at the output level, it does not treat the groups of individuals differently depending on their membership of protected categories. Instead, it applies the same treatment to all decision-subjects. This is beneficial because, given that the law may sometimes prohibit affirmative action, this method assists in keeping within these legal boundaries.

Disparate treatment is built on the idea of neutrality and non-classification: treating men and women differently is not permitted because decision-makers must be neutral about gender.[225] This information approach does not alter the process of the decision. Because it addresses discrimination by dealing with the input data not the decision process, it separates the data problem (bias of input

220. *Id.*

221. *Id*. In other words, demographic parity as a fairness criterion works without doing affirmative action (and thereby decreasing prediction accuracy) only when one assumes that there are no intrinsic differences between the groups. *Cf.* Chander, *supra* note 39 (proposing a system of algorithmic affirmative action).

222. John Lightbourne, *Damned Lies and Criminal Sentencing Using Evidence-Based Tools*, 15 DUKE L. & TECH. REV. 327, 337–342 (arguing this for algorithmic risk assessment, where the Equal Protection clause applies); Starr, *supra* note 39, at 827.

223. Barocas & Selbst, *supra* note 21, at 724-28; Kroll et al., *supra* note 15 at 692–94. *Cf. Data Driven Discrimination*, *supra* note 21, at 925–932 (arguing instead that *Ricci* would not apply to this context).

224. *See generally* Muhammad Bilal Zafar et al., *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment*, *in* 2017 INT'L CONF. ON WORLD WIDE WEB 1171 (2017), https://doi.org/10.1145/3038912.3052660 (developing the concept of disparate mistreatment, defined in terms of misclassification rates, to avoid different misclassification rates across groups at a small accuracy cost).

225. *See generally* Paul Brest, *Forward: In Defense of the Antidiscrimination Principle*, 90 HARV. L. REV. 1 (1976) (advocating for the continued use of the antidiscrimination principle).

data) from the algorithmic problem (measurement bias).[226] Therefore, it would not fall under the constitutional challenges based on disparate treatment.[227] This illustrates a benefit of the information approach to discrimination presented here. Preventing algorithmic discrimination through an information policy addresses disparate impact concerns with a disparate treatment logic. Therefore, it can address discrimination in terms of antisubordination in those areas of the law where disparate impact is not recognized.

Information rules are most useful in cases in which disparate impact doctrine cannot be applied, leaving groups under-protected by traditional antidiscrimination law.[228] That is, when there is no statute explicitly incorporating disparate impact as Title VII and the Federal Housing Act do. Many decisions in daily life, while arguably not as crucial as employment or housing, significantly affect people's quality of life and depend on decision-makers who might unintentionally discriminate against protected categories.[229] For example, some argue that men tend to mentor more men than women—a phenomenon that should raise concern especially because more men than women occupy positions of power.[230] There is also evidence that doctors tend to provide pain medication to white patients more than to black patients.[231] Most notably, while the the Consumer Financial Protection Bureau considers that disparate impact applies to ECOA,[232] this may change in the future as there is yet no Supreme-Court-recognized disparate impact protection for decisions on loan applications, which significantly affect people's ability to obtain housing.

Information rules will also be useful when a disparate impact statute applies but there are vulnerable groups that are not protected by them. These rules can be used, for example, to protect LGBTQ individuals from employment

226. *See, e.g.*, Johndrow and Lum, *supra* note 195 (arguing that one should separate the data problem from the process problem and develop an algorithm that makes one variable in the input set independent of the outcome of the model so that, instead of removing the variable, they propose a method to create outputs that are independent of the "protected" variable).

227. *See* Primus, *supra* note 210, at 494–95 (showing that disparate impact standards such as Title VII are not unconstitutional, but tensions exist between these standards and a disparate treatment view of the Equal Protection Clause). Current affirmative action cases are good examples of constitutional challenges under a disparate treatment theory. *See generally* Fisher v. Univ. of Tex. at Austin, 136 S. Ct. 2198 (2016) (holding that a university's "race conscious" admission process did not violate equal protection); Students for Fair Admissions, v. Harvard College, 261 F. Supp.3d 99 (D. Mass. 2017) (considering whether race was considered a "'plus factor'" in favor of admission to the school).

228. *Cf.* Roberts, *supra* note 12, at 2123–34 (stating that privacy belongs to the realm of anticlassification and antisubordination requires providing more information).

229. *See generally* Colleen Sheppard, *Institutional Inequality and the Dynamics of Courage*, 31 WINDSOR YEARBOOK OF ACCESS TO JUSTICE 103 (2013) (discussing institutionalized inequality and systemic discrimination, and arguing that retroactive legal remedies are ineffective at addressing them).

230. *See, e.g.*, KIM ELSESSER, SEX AND THE OFFICE: WOMEN, MEN, AND THE SEX PARTITION THAT'S DIVIDING THE WORKPLACE 4 (2015).

231. Kelly M. Hoffman et al., *Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs About Biological Differences Between Blacks and Whites*, 113 PROC. NAT'L ACAD. SCI. 4296, 4296 (2016); Sophie Trawalter et al., *Racial Bias in Perceptions of Others' Pain*, PLOS ONE, Nov. 2012, at 1 (2012).

232. CONSUMER FIN. PROT. BUREAU, CFPB BULLETIN 2012-14 (FAIR LENDING) (2012) https://files.consumerfinance.gov/f/201404_cfpb_bulletin_lending_discrimination.pdf.

discrimination, given that a number courts have held that Title VII protects discrimination based on gender but not sexual orientation.[233] In the same vein, privacy rules can be used for any effort to protect other minorities not considered a protected category under Title VII, such as resident legal aliens.[234]

In sum, this information approach can expand protection from discrimination beyond what standard measures of antidiscrimination can do. Moreover, it widens the scope of antidiscrimination protection. Besides doing so for those cases in which disparate impact is not recognized,[235] it can be used when disparate impact is recognized but there are probatory difficulties involved.[236] These probatory difficulties exist, for example, when disparate impact itself is difficult to prove (such as in housing) or when a party raises the business necessity defense and its discriminatory intent is difficult to prove.

B. THE ANTISUBORDINATION PRINCIPLE IN ALGORITHMIC DISCRIMINATION

Reva Siegel and Jack Balkin have articulated the understanding of antidiscrimination as falling under two competing principles: anticlassification and antisubordination.[237] Anticlassification prohibits classifying based on protected categories like gender or race. Antisubordination prohibits disadvantaging or aggravating historically vulnerable groups like women or Latinos.[238]

The logics of anticlassification and antisubordination overlap with the distinction between disparate treatment and disparate impact in antidiscrimination law.[239] Disparate impact can be used either as its own

---

233. *See, e.g.*, Boy Scouts of America v. Dale, 530 U.S. 640 (2000) (holding that forcing the Boy Scouts to include a homosexual man violates their First Amendment freedom to express that homosexuality is inappropriate); DeSantis v. Pacific Tel. & Tel. Co., 608 F.2d 327 (9th Cir. 1979) (holding that sexual identity is not covered by Title VII, and emphasizing that that Congress chose not to pass amendments to extend Title VII to cover sexual preferences). *But see* Rene v. MGM Grand Hotel, 305 F.3d 1061 (9th Cir. 2002) (extending protection because the claim included unwanted physical conduct, considered by the court to be always of a sexual nature and therefore sex discrimination).

234. *See, e.g.*, Espinoza v. Farah Mfg. Co., 414 U.S. 86 (1973) (holding that Title VII protection on national origin does not extend to alienage or citizenship).

235. *See supra* Subpart IV.A.

236. *See Protecting Privacy*, *supra* note 12, at 2149–55.

237. *See* Bornstein, *supra* note 62, at 540–43 (applying anticlassification and antisubordination to algorithmic discrimination in employment). *See generally* Jack M. Balkin & Reva B. Siegel, *American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9 (2003); Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over* Brown, 117 HARV. L. REV. 1470 (2004).

238. Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 108, 157 (1976).

239. Balkin & Siegel, *supra* note 237, at 12 ("Fiss and the audience of *Groups and the Equal Protection Clause* understood the anticlassification and antisubordination principles to have divergent practical implications for key issues of the moment: The anticlassification principle impugned affirmative action, while legitimating facially neutral practices with racially disparate impact, while the antisubordination principle impugned facially neutral practices with a racially disparate impact, while legitimizing affirmative action."); *see also* Susan D. Carle, *A Social Movement History of Title VII Disparate Impact Analysis*, 63 FLA. L. REV. 251 (2011); Ruth Colker, *Anti-Subordination Above All: Sex, Race, and Equal Protection*, 61 N.Y.U. L. REV. 1003 (1986). *But see* Ian Ayres & Peter Siegelman, *Q-Word as Red Herring: Why Disparate Impact Liability Does Not Induce*

standard under the antisubordination principle or as mere evidence for disparate treatment under anticlassification. The standard view is that American law predominantly follows the anticlassification approach (race or gender blindness) with the exception of a few statutes with an antisubordination orientation, such as Title VII;[240] but other historical accounts have stated that courts have shifted ambivalently between both principles. [241]

Applying ex-post antidiscrimination law based on an anticlassification approach to algorithms will not avoid their discriminatory outcomes when they are based on biases in the sample data or societal biases embedded in representative data.[242] For example, consider again the Amazon hiring algorithm. After realizing that the algorithm discriminated based on gender, Amazon modified it to ignore words that denoted gender. However, the algorithm continued to "guess" individuals' gender by using other words in the resume that correlated with gender. Courts would not understand this as going against the anticlassification principle, but would rather consider it to be a case of disparate impact, because classifying based on proxies for protected categories is not disparate treatment. If an anticlassification approach were applied in a lawsuit against the company, therefore, this bias would remain.

However, one does not need a notion of antisubordination to modify an algorithm's training data (which, in some way, operates as blindness). Being operable under an anticlassification paradigm makes antidiscriminatoty information rules compatible with the mainstream of antidiscrimination doctrine and case law. This is an advantage because it makes this proposal viable under anticlassification-dominated doctrine, even outside of the scope of statutes that recognize disparate impact such as Title VII.[243]

At the same time pre-processing data in such way is compatible with the logic of antisubordination. As Owen Fiss shows, the purpose of the antisubordination principle, underlying in disparate impact discrimination, is that decisions should not worsen *or* perpetuate protected groups' subordinate status.[244] The purpose of disparate impact-focused antidiscrimination is not to

Hiring Quotas 74 TEX. L. REV. 1487, 1489 (1996) ("[F]ar from producing hiring quotas that induce employers to discriminate in favor of minorities, disparate impact liability may actually induce hiring discrimination against minorities (and other protected groups).").

240. Bradley A. Areheart, *The Anticlassification Turn in Employment Discrimination Law*, 63 ALA. L. REV. 955, 960–67 (2012); *see also* Catherine A. MacKinnon, *Unthinking ERA Thinking*, 54 U. CHI. L. REV. 759, 765 (1987) (proposing that the liberal interpretation of ERA mistakenly reduced "the problem of the subordination of women to men to a problem of gender classification . . . .").

241. *See generally* Siegel, *supra* note 237 (detailing how the courts tend to fluctuate between the two principles); Reva Siegel, *From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases*, 120 YALE L.J. 1278 (2011) (tracing a third understanding, in between these two, under which equal protection strives not to achieve colorblindness or protection from subordinating practices, but protection from the threat of society's balkanization; courts concerned with antibalkanization focus on diversity more than on equality and that some antisubordination-based strategies might generate further divisions in society).

242. Corbett-Davies and Goel, *supra* note 67, at 3.

243. 42 U.S.C. § 2000 (2012).

244. *See* Fiss, *supra* note 238, at 157; *see also* Cass R. Sunstein, *The Anticaste Principle*, 92 MICH. L. REV. 2410, 2415–16 (1994).

freeze inequality in an unjust situation but rather to correct the state of affairs and achieve a more just state of the world.[245] This is, as discussed above, what training algorithms with fair data would accomplish.

This builds on the idea discussed earlier that what is needed is not more data, but more meaningful data.[246] More meaningful data, counterintuitively, means a data sample that is unrepresentative of the pool, because it looks like what we believe the pool *would* look like had it not embedded structural inequalities. The objection of the endless line of proxies indeed raises an important concern. If one cares about disparate impact, protected attributes must be included in the data, but they must be included with the objective of training the system in a way that avoids disparate impact.[247]

To follow an anticlassification approach is to behave like Amazon did after discovering the effect of its hiring algorithm: blocking the protected category. If disparate impact is applied but merely as evidence for disparate treatment, blocking the protected category (in this case, gender) is sufficient to avoid disparate treatment because blindness guarantees a lack of differential treatment among groups.[248] But, because of the endless line of proxies described above,[249] blindness will not avoid disparate impact outcomes. To avoid such outcomes, one must do something different than either blocking or allowing information about protected categories. To apply the antisubordination principle to algorithmic discrimination is to lie to the machine. It is to represent to the machine the world that we want as opposed to replicating the world that we have.

C. AN ALTERNATIVE TO THE ALGORITHMIC FAIRNESS IMPOSSIBILITY

Discrimination discovery through concepts of algorithmic fairness, like disparate impact discrimination, are standards that the algorithmic decision must satisfy ex-post.[250] The focus of this Article is not to elucidate such standards, but to explore how they can be achieved ex-ante.

---

245. *See* Fiss, *supra* note 238, at 176 (distinguishing between equal treatment and equal status); *see also* Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67, 69 (2017).

246. *See* Barocas and Selbst, *supra* note 105 (arguing that what is needed is not more data, but meaningful data); *see also supra* Subarts II.B and II.C.

247. Cynthia Dwork et al., *Decoupled Classifiers for Group-Fair and Efficient Machine Learning*, *in* CONF. ON FAIRNESS, ACCOUNTABILITY AND TRANSPARENCY 119, 120 (2018), http://proceedings.mlr.press/v81/dwork18a.html (providing an example of disparate learning process and suggesting training each group separately).

248. *See* Balkin & Siegel, *supra* note 237,at 9–11; Colker, *supra* note 239, at 1005–06.

249. *See supra* Subpart II.D.

250. *See generally* Žliobaitė, *supra* note 45, at 5 ("Discrimination prevention develops machine learning algorithms that would produce predictive models, ensuring that those models are free from discrimination, while, standard predictive models, induced by machine learning and data mining algorithms, may discriminate groups of people due to training data being biased, incomplete, or recording past discriminatory decisions.").

Recall the COMPAS recidivism algorithm and the Amazon employment algorithm. *The problem with these algorithms is not that they are biased, but rather that they reinforce the world that we have*.[251]

To determine whether such processes are discriminatory, computer scientists apply mathematical notions of fairness.[252] While there are many such notions, they can broadly be categorized into group fairness criteria and individual fairness criteria.[253] Individual criteria are based on predicted criterion scores for each individual, such as an equal rate of false positives and false negatives for individuals in each group. [254] Group criteria are based on the distribution of criterion scores between the population groups, such as the protected category having an equal proportion of positive classification than the overall population.[255]

The problem with "fixing" these algorithms is that one cannot have calibration between groups (group fairness) at the same time that one controls for individual classification (individual fairness).[256] In other words, calibration and equal false positives and false negative rates for individuals cannot be satisfied at the same time.[257] Making a test that has both (i) the same number of false positives and false negatives for individuals across populations and (ii) the same level of accuracy among populations is only possible when the populations are identical for the purposes of the analysis.[258]

We can see this dichotomy in the tension between two concepts of algorithmic fairness: predictive accuracy and statistical parity. Predictive accuracy measures one group of individuals against another: the error rates in classification (false positives and false negatives) should be the same for both groups.[259] Statistical parity measures positive predictions against all predictions, and poses that the rate of positive predictions should be the same across groups:

---

251. This is the decision-making problem in which the algorithm perpetuates and amplifies societal biases. *See supra* Subpart II.D.

252. *See generally* Ben Hutchinson & Margaret Mitchell, *50 Years of Test (Un)fairness: Lessons for Machine Learning*, *in* 2019 CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY (2018), https://arxiv.org/pdf/1811.10104.pdf.

253. *See* Zemel et al., *supra* note 181, at 325.

254. *See, e.g.*, Chouldechova, *supra* note 38, at 133. *See generally* Hardt et al., *supra* note 167.

255. Dwork et al., *supra* note 145, at 18 (proposing an individual fairness framework based on a task-specific and externally defined similarity metrics between individuals, under the principle that "similar people [should be] treated similarly"). *See generally* Sorelle A. Friedler et al., *On the (Im)possibility of Fairness*, (Sept. 23, 2016) (unpublished manuscript), https://arxiv.org/pdf/1609.07236.pdf; Zemel et al., *supra* note 181.

256. Jon Kleinberg, *Inherent Trade-Offs in Algorithmic Fairness*, *in* ABSTRACTS OF THE 2018 ACM INT'LCONF. ON MEASUREMENT AND MODELING OF COMPUTER SYS. 40, 43–44 (2018), http://doi.acm.org/10.1145/3219617.3219634.

257. Calibration requires that the expected proportion of individuals for each group that receives a positive (or negative) outcome are equivalent. *Id.*

258. *See* FRY, *supra* note 2, at 66–69 (explaining the intuition behind this and providing a numerical example).

259. Chouldechova, *supra* note 38, at 157 ("[I]f an instrument satisfies predictive parity . . . but the prevalence differs between groups, the instrument cannot achieve equal [false positives] and [false negatives] across those groups."); *see also* Corbett-Davies, *supra* note 151; Corbett-Davies et al., *supra* note 177; Hardt et al., *supra* note 167, at 19.

both groups should have equal fractions labeled as positive.[260] While predictive accuracy is an individual fairness criterion, statistical parity is a group fairness criterion,[261] and one cannot achieve individual notions of fairness and group notions of fairness at the same time.[262] The first criterion requires the algorithm to be as likely to be wrong about whether I will recidivate, or be a good employee, as it will with someone from the other group.[263] The second criterion requires the algorithm to predict that an equal proportion of both groups (e.g. of white and black individuals) will recidivate, or be good employees.[264] However, if the two populations in fact have different recidivism rates, or different employee satisfaction scores, then the algorithm cannot simultaneously be consistently accurate with all individuals and predict equal levels of success for both groups.

In the COMPAS example, this means that for Northpointe to develop an algorithm with equal predictive power for white defendants and black defendants given the data that Northpointe had, the algorithm would necessarily have different rates of false positives and false negatives for both types of defendants.[265] This holds even without collecting information about race.[266]

The broader issue, in other words, is that some notions of algorithmic fairness are incompatible with others.[267] This means that a normative choice is

260. *See* Feldman et al., *supra* note 186, at 261–63 (explaining statistical parity); *see also* Calders & Verwer, *supra* note 181, at 285-290; Kamishima et al., *supra* note 142, at 643.

261. MacCarthy, *supra* note 245, at 102; *see also* Chouldechova, *supra* note 38, at 157 ("[I]f an instrument satisfies predictive parity . . . but the prevalence differs between groups, the instrument cannot achieve equal [false positives] and [false negatives] across those groups."); Hutchinson & Mitchell, *supra* note 252 (explaining how most modern machine learning notions of fairness map to mathematical notions of fairness of the 1970s and 1980s in the fields of education and employment).

262. Friedler et al., *supra* note 255; *see also* Richard L. Sawyer, Nancy S. Cole & James W. L. Cole, *Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection*, 13 J. EDUC. MEAS. 59, 69 (1976) ("[M]aximization procedures based on individual parity do not produce equal opportunity (equal selection for equal success) based on group parity and the opportunity procedures do not produce success maximization (equal treatment for equal prediction) based on individual parity.").

263. Zemel et al., *supra* note 181, at 325 (defining individual fairness as "similar individuals should be treated similarly" [independent of membership to a protected category]); *see also* Dwork et al., *supra* note 145, at 2 (proposing such definition).

264. Zemel et al., *supra* note 181, at 325 (defining group fairness as "the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole"). *See generally* MacCarthy, *supra* note 246 (defending the use of group fairness in algorithmic design).

265. *See generally* KHALIL GIBRAN MUHAMMAD, THE CONDEMNATION OF BLACKNESS (2010) (exploring racial bias in crime data). Northpointe repeatedly claimed, in fact, that its scores were well calibrated, as unconvincing as one could find this claim. *See* Adam Liptak, *Sent to Prison by a Software Program's Secret Altorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html.

266. *See* Corbett-Davies et al., *supra* note 177 (analyzing COMPAS in terms of predictive accuracy and statistical parity).

267. Friedler et al., *supra* note 255 (showing the tension between individual and group notions of fairness). *See generally* Sawyer et al., *supra* note 262. The problem of applying a machine learning algorithm to a different context, which may require a different definition of fairness, has been called the portability trap. Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, *in* 2019 CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 59, 61 (2019), http://doi.acm.org/10.1145/3287560.3287598.

necessary. One can either define fairness in an individual way (in terms of false positives and false negatives) or one can define fairness in terms of equal predictive accuracy across groups.[268] This takes place because, like any other risk estimates, fairness criteria will be incompatible when base rates among groups are not the same.[269] We simply cannot have both at the same time with non-identical groups.[270] That is, unless we alter the data about them.

While a sizeable amount of research on algorithmic fairness focuses on the relationship between concepts of fairness and tradeoffs between fairness and accuracy,[271] recent research has also pointed out that, when unfairness is introduced by small sample sizes or unmeasured predictive variables, it may be more effective to address the problem at the data collection stage.[272] Similarly, several industry members with research divisions that focus on algorithmic bias, such as IBM and Microsoft, have emphasized the issue that machine learning algorithms are only as good as the data that we feed them.[273]

While this research focuses on collecting *more* data, which the legal literature has already problematized,[274] we can take their considerations to argue instead for *different* data.

D. AVOIDING ALGORITHMIC OPACITY

One of the central issues with algorithmic decision-making is its opacity problem. A spam filter, for example, uses classifiers and predictors to determine whether an email is likely spam, but it cannot explain why it is such. Credit card fraud detection algorithms follow the same dynamic, as do credit scoring and loan decision algorithms.[275] In addition, the most accurate methods of machine

268. *See, e.g.*, Corbett-Davies, *supra* note 151.

269. Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, *in* 2017 PROC. IN INNOVATIONS & THEORETICAL COMPUTER SCI. 1, 17 (2017).

270. Geoff Pleiss et al., *On Fairness and Calibration*, *in* 2017 CONF. ON NEURAL INFO. PROCESSING SYS. 1, 8 (2017), http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf; *see also* Žliobaitė, *supra* note 45, at 3 ("Quite often research papers propose a new way to quantify discrimination, and a new algorithm that would optimize that measure. The variety of approaches to evaluation makes it difficult to compare the results and assess the progress in the discipline, and even more importantly, it makes it difficult to recommend computational strategies for practitioners and policy makers.").

271. Feldman et al., *supra* note 181, at 3–4; Indre Žliobaitė, *On the Relation Between Accuracy and Fairness in Binary Classification*, *in* 2015 WORKSHOP ON FAIRNESS, ACCOUNTABILITY, TRANSPARENCY IN MACHINE LEARNING (2015), http://arxiv.org/abs/1505.05723.

272. Irene Chen et al., *Why Is My Classifier Discriminatory?*, *in* 2018 CONF. ON NEURAL INFO. PROCESSING SYS. (2018), http://arxiv.org/abs/1805.12002 .

273. *See* Ruchir Puri, *Mitigating Bias in AI Models*, IBM RES. BLOG (Feb. 6, 2018), https://www.ibm.com/blogs/research/2018/02/mitigating-bias-ai-models; John Roach, *Microsoft Improves Facial Recognition to Perform Well Across All Skin Tones, Genders*, MICROSOFT: AI BLOG (June 26, 2018), https://blogs.microsoft.com/ai/gender-skin-tone-facial-recognition-improvement/.

274. *See generally* Barocas & Selbst, *supra* note 21.

275. Danielle Keats Citron, *Technological Due Process*, 85 WASH. UNIV. L. REV. 1249, 1286–87, 1289 (2008). *See generally* PASQUALE, *supra* note 1; Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y, Jan. 6, 2016 at 1.

learning, and thus the ones for which there is greater incentive to adopt, seem to be the least explainable ones.[276]

Opacity introduces a regulatory problem not only because decision-subjects may have a right to an explanation,[277] but also because it makes it more difficult to create an environment that reduces the existence of biases, [278] and it limits the application of doctrines such as disparate impact. [279] It is difficult, in other words, to correct a decision-making process that we cannot access or understand.[280] While this Article is not concerned with explainability and procedural rights, opacity is relevant for the consequences of algorithmic decisions because evaluating such consequences becomes difficult where the algorithm is opaque.[281]

One should be note that opacity, or inscrutability, is not unique to automated decision-making—humans can be black boxes, too. Many human decisions seem inscrutable and opaque.[282] One reason for which opacity is still relevant in algorithmic decision-making is that the law is used to treating humans as a black box, but it does not always do so with algorithms.[283] Another reason for which opacity and inscrutability are relevant is because humans can choose to deploy algorithms in environments, such as hiring, where their opacity may be strategically advantageous to human decision-makers who could have chosen a different system because it adds an additional layer of obfuscation to access

276. Solon Barocas, *Understanding Inscrutability*, ALGORITHMS EXPLAN. CONF. PAP. (2018).

277. Opacity refers to the lack of understanding of how a decision-making algorithm arrives at its outputs from its inputs. Burrell, *supra* note 275, at 2.

278. Miriam C. Buiten, *Towards Intelligent Regulation of Artificial Intelligence*, 10 EUROPEAN J. OF RISK REG. 41, 43 (2019).

279. *See Data-Driven Discrimination*, *supra* note 21; *see also Big Data and Artificial Intelligence*, *supra* note 21, at 12 ("Another challenge is that the factors driving the results may be unknown. An algorithm can be so complex that its decision process is completely opaque—even to the programmers who created it. . . . Given these characteristics, biased algorithms raise many questions and challenges for disparate impact doctrine. If an algorithm produces a racially discriminatory effect, can the employer meet its burden of showing that it is 'job related' by demonstrating that it rests on a robust statistical correlation? Even if the correlation is unexplained?").

280. Citron & Pasquale, *supra* note 35, at 18–20 (arguing that procedural regularity is essential for those stigmatized by AI scoring systems, and the U.S. tradition of due process should inform basic safeguards; the law should open algorithmic black boxes and allow people to examine them). *See generally* Citron, *supra* note 275 (highlighting accountability deficits and arguing that a new concept of technological due process essential to uphold procedural protections).

281. *See* Cynthia Rudin et al., *The* Age *of Secrecy and Unfairness in Recidivism Prediction* (Nov. 2 2018) (unpublished manuscript), http://arxiv.org/abs/1811.00731 (reverse engineering the COMPAS algorithm and using it as an example of how lack of transparency leads to uncertainty as to whether an algorithm meets any standard of fairness); *Big Data and Artificial Intelligence*, *supra* note 21, at 13 ("Existing disparate impact doctrine isn't equipped to deal with issues like these, and so to address classification bias, the law needs to recognize that predictive algorithms differ from traditional ability tests and to adapt accordingly.").

282. We humans are able to produce the equivalent of algorithmic post-hoc explainability through verbal accounts of a decision process. This can be done in deep learning algorithms, for example, through saliency maps. But requiring post-hoc explainability despite a black box nature and requiring interpretability (or intelligibility) are different demands. *See* Lipton, *supra* note 53; *see also* Kroll et al., *supra* note 15, at 657-59 (arguing against the idea that transparency is a panacea).

283. Katherine Strandburg, *Algorithmic Explainability* (2019) (on file with author); Cofone & Strandburg, *supra* note 130.

people's true motivations.[284] A third reason for which this is relevant is that, while human opacity is inevitable—and the best thing we can do is demand good-faith explanations—algorithmic opacity is not.[285] In terms of regulating information, algorithms that may discriminate place regulators in a better position than they were before, as they provide an opportunity to regulate the data in ways that one cannot do for human decision-makers.

Opacity is particularly problematic when it obfuscates intent. Discriminatory intent, while required to establish disparate treatment discrimination,[286] is often difficult to prove in the context of algorithmic decision-making.[287]

By focusing on the acquisition of suspected information rather than on its use, and by depriving decision-makers of these information points, antidiscriminatory information rules avoid the need for proving what is in the mind of the employer.[288] This proof is difficult to obtain as employers have incentives to hide discriminatory intent and can do so especially when they have mixed motives.[289]Information rules, in other words, are effective at dealing with facially neutral screening rules set with ulterior motives.[290]

Opacity has been classified in three types.[291] The first is intentional opacity, when the process is deliberately hidden, as in trade secrets. For example, COMPAS, described above, does not ask for race as an input. However, some of its data points, which have different weights, correlate to race. It is difficult to know exactly how this took place because the algorithm is a trade secret, so its code is unavailable to the public and to those who are granted or refused parole based on it.[292] The second is opacity as result of some inevitable degree

284. Nicholas Diakopoulos, *Accountability in Algorithmic Decision Making*, 59 COMMC'NS OF THE ASS'N. OF COMPUTING MACHINERY 56, 59–61 (2016).

285. Depending on the algorithm, it can be harder to reverse-engineer human decisions than it is fto reverse-engineer algorithmic decisions. I explore different types of algorithms in terms of their opacity in Part IV.

286. *See* Washington v. Davis, 426 U.S. 229, 239–41 (1976) (holding a showing of discriminatory intent may violate Equal Protection Clause).

287. Kim & Scott, *supra* note 51, at 24 ("Disparate treatment cases turn on employer intent, and thus, whether an employer's online targeting strategy supports a finding of liability depends upon how clearly it indicates a discriminatory preference. If the employer expressly excludes some social media users from its target audience because of their protected characteristics, those choices strongly suggest that it intends to discourage members of those groups from applying. For example, if an employer directs its advertising only at men, or only at persons aged 18 to 35, a court may infer that a female or older applicant was rejected because of the employer's discriminatory motive. Less explicit strategies, such as selecting an audience using neutral attributes or relying on the lookalike audience tool, may not clearly indicate a discriminatory preference, making it more difficult to infer motive from these choices.").

288. *See* Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 991 (1988).

289. *See* Price Waterhouse v. Hopkins, 490 U.S. 228, 247 n.12 (1989); Fuller v. Phipps, 67 F.3d 1137, 1141-42 (4th Cir. 1995); Tyler v. Bethlehem Steel Corp., 958 F.2d 1176, 1181 (2d Cir. 1992); *see also* Desert Palace v. Costa, 539 U.S. 90, 101–02 (2003) (holding that overt discrimination is not always required in mixed motive cases).

290. Cofone, *supra* note 11, at 162–64.

291. Burrell, *supra* note 275, at 1–2.

292. *See* Cofone & Strandburg, *supra* note 130 (exploring when subjects of algorithms can and should be told how the decision process works while avoiding gaming concerns).

of the general public's technical illiteracy: it would be too costly to teach complex processes to all the population. The third is opacity as a result of the characteristics and scale of an algorithm, such as deep learning algorithms, for which a certain degree of opacity is inherent to the learning process.[293]

The three kinds of opacity require different kinds of responses if the outcome of the algorithm being used is deemed discriminatory. For the first, disclosure of the decision-making process can be mandated, especially to determine whether antidiscrimination law is being disregarded.[294] For the second, general education to understand the process can be put in place or expert auditing can be implemented (leading to the question of for whom should the algorithm be made transparent).[295]

For the third, opacity that is inherent to a decision-making algorithm, determining the type of response required is more difficult. Deep learning algorithms, for example, lead to the third type of opacity because, with them, we simply have raw data in the world and a black box that processes such data,[296] so it is difficult to know how discriminatory outcomes are reached. However, because as discussed above, humans are a black box too, the law has experience in addressing that. For deep learning algorithms, then, we may draw upon the principles that we have for humans. Counterintuitively, the most advanced technology is the technology that may turn out to be least disruptive for the law.[297]

Opacity generates another related problem for algorithmic bias. Opacity makes liability difficult to determine: harmed parties need some level of transparency in order to substantiate their claims.[298] Transparency in terms of interpretability is imporant for plantiffs because it is key for determining whether the model complied with technical and legal standards. But definitions

293. Burrell, *supra* note 275, at 1–2.

294. *See generally* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018) (arguing that algorithms under trade secret laws should disclose their decision-making process when being used in criminal procedure); Rebecca Wexler, Opinion, *When a Computer Program Keeps You in Jail*, N.Y. TIMES (June 13, 2017), https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html.

295. *See Big Data and Artificial Intelligence*, *supra* note 21, at 13 (arguing that, despite its limitations, auditing for discrimination should remain an important part of the strategy to detect and respond to bias in algorithms); Roth, *supra* note 107, at 1984–85 (questioning the role of algorithms in trials and arguing that, in evidence law, more than accuracy, our worry should be whether the jury has enough tools to interpret the algorithm).

296. *See* Leilani H. Gilpin et al., *Explaining Explanations: An Overview of Interpretability of Machine Learning*, *in* 2018 IEEE INT'L CONF. ON DATA SCIENCE & ADVANCED ANALYTICS (2018), https://arxiv.org/pdf/1806.00069.pdf ("[T]he fundamental problem facing explanations of such processing is to find ways to reduce the complexity of all these operations . . . . One common viewpoint in the deep neural network community is that the level of interpretability and theoretical understanding needed to for transparent explanations of large DNNs remains out of reach."); *see also supra* Subpart IV.A.

297. Cofone, *supra* note 11(arguing that end technologies are sometimes the least disruptive because they can be analogized to humans, and transition technologies are sometimes the most disruptive because they don't share features that the law picks up neither with humans nor with objects).

298. Buiten, *supra* note 278, at 57 (adding that this obstacle may also create difficulties in the protection of fundamental rights); *see also* Rudin et al., *supra* note 281, at 2–3.

of transparency are diverse, they present tradeoffs with each other, and they are not always achievable.[299]

Antidiscriminatory information rules have the advantage that, because they focus on regulating the data rather than on regulating the algorithm, they function on opaque models as well as on interpretable ones. While a black box algorithm can contain discrimination that is invisible, making it impossible to identify and mitigate the causes of its discriminatory outcome, it is possible to shape and encode the training data of a black box algorithm.

Thus, the antidiscriminatory information rules explored here are particularly relevant when ex-post regulation or liability are faced with obstacles related to algorithmic opacity. Because of opacity, applying traditional, ex-post antidiscrimination law is significantly more difficult for black box algorithms than it is for humans; however, because one can regulate their input data, applying an information policy is actually easier for black box algorithms than it is for human decision-makers.

E. THE VALUE AND COST OF EX-ANTE REGULATION

Implementing a preventive approach that addresses discrimination through information is particularly valuable in the context of algorithms. This is not only because of the disparate-impact disparate-treatment tension, which leads antidiscrimination law to be ineffective at dealing with algorithms,[300] and the insufficient coverage of disparate impact in American law. It is also valuable for alogorithms from a policy perspective because, for victims of discrimination, preventing discriminatory harm is more valuable than repairing it.

At a policy level, by operating ex-ante, this approach avoids the social harms created by discriminatory conduct, which are only partially solved by ex-post compensation.[301] The harm produced by discrimination is only partly monetary, and even to the extent that it is monetary it is extremely difficult to compensate fully.[302] While antidiscrimination litigation is of crucial importance to providing redress to those who have suffered discriminatory harm and to discourage discrimination *ex ante*, developing preventive measures that aim to eradicate discrimination more directly is of crucial importance.[303] This is particularly the case for addressing systemic discrimination, which is difficult to compensate and detect.[304]

---

299. Lipton, *supra* note 53, at 98–99.

300. *See* Barocas & Selbst, *supra* note 21, at 694–712.

301. *See* Austin, *supra* note 98, at 41–55 (presenting a discussion of privacy acting as a preventive or anticipatory remedy).

302. *See Protecting Privacy*, *supra* note 12, at 2155–56.

303. *See* MARIE MERCAT-BRUNS, DISCRIMINATION AT WORK: COMPARING EUROPEAN, FRENCH, AND AMERICAN LAW 108 (2016) ("Although litigation is important for bringing to light purportedly objective requirements perpetuating workforce segregation, prevention is key to eliminating systemic discrimination.").

304. *See generally Data-Driven Discrimination*, *supra* note 21 (explaining that the consequences of algorithmic discrimination take place on a large scale rather than an individual scale); Sheppard, *supra* note 229 (discussing the challenges of systemic discrimination for antidiscrimination law).

This regulation of information, however, comes at a cost. This type of regulation would often imply a loss in accuracy with respect to the output variable. If, continuing with the examples used, "re-arrest" is taken as an output variable to gauge recidivism or "current employees" is taken as an output variable to gauge good job candidates, altering the input data in such a way will reduce the accuracy of the model to predict the output variable, even if it may or may not do so for what the model is actually supposed to predict. [305] In other words, these algorithms would have lower predictive accuracy with regards to their human-determined *proxy*—similarity with current employees and likelihood of re-arrest—but not necessarily with regards to the target variable.

If one believes that the Amazon algorithm and COMPAS were biased with respect to the target variable (desirableness as an employee or likelihood to recidivate), this should lead one to believe that, when trained with fair data, they would not have lower predictive accuracy with respect to the desired prediction.[306] The adjustment would correct for the error between the chosen proxy and the target variable: between similarity with current employees and likelihood of re-arrest and desirableness as an employee and likelihood of re-offense.

But imagine one held the (arguably questionable) belief that the algorithms were not disadvantaging a protected category by hiring fewer women and imprisoning more black individuals, and therefore that training them with fair data would lead to lower predictive accuracy with respect to the target variable. Whether one considers the algorithms as they exist as discriminatory would depend on whether there are duties of reasonable accommodation, which hinge upon the antidiscrimination principle used.[307] For example, courts have ruled in the past that no-beards policies disproportionately affect black men and therefore constitute disparate impact discrimination, even if this implies a minor economic loss and therefore some duty of accommodation for the employer.[308]

Under certain circumstances, the law will consider a decision as discriminatory even if the prediction is true because what is required under disparate impact is for the employer, or court, to adjust its expectation or provide options for the person to perform the task.[309] Because algorithmic discrimination is mainly a disparate impact problem,[310] when disparate impact is identified one

305. *But see* Feldman et al., *supra* note 181 (developing a method to remove information on protected category without reducing accuracy by preserving each individual's rank orthogonal to their class membership and showing that, while removing gender in a database fails at preventing discrimination, a number of pre-processing techniques eliminates discrimination with a minimal loss in accuracy).

306. *See* Lehr & Ohm, *supra* note 17, at 704 (noting that overfitting generates less accurate predictions for minority groups).

307. *See supra* Subparts IV.A., IV.B.

308. *See, e.g.*, Bradley v. Pizzaco of Nebraska, 7 F.3d 795, 796 (8th Cir. 1993); Richardson v. Quik Trip Corp., 591 F. Supp. 1151, 1155-56 (S.D. Iowa 1984); EEOC v. Trailways, Inc., 530 F. Supp. 54, 59 (D. Colo. 1981). *But see* Equal Emp. Opportunity Comm'n v. Greyhound Lines, 635 F.2d 188, 189 (3d Cir. 1980); Lewis v. Univ. of Pa, No. 16 Civ. 5874, slip. op. (E.D. Pa. Jan. 29, 2018).

309. *See* Christine Jolls, *Antidiscrimination and Accommodation*, 115 HARV. L. REV. 642, 697–99 (2001).

310. *See generally* Barocas & Selbst, *supra* note 21.

should expect the law to introduce some cost, such as reduced overall accuracy, in a similar way that disparate impact antidiscrimination law does for human decisions.[311] While the cost of poor classification is currently borne by minorties and other vulnerable groups, shifting this cost to decision-makers would provide incentives to invest in improving classification accuracy.[312] One could therefore establish that, when there is a discriminatory outcome, the creator of a decision-making algorithm has a burden of proof on adequacy of data.

The fact that a regulation brings some level of costs does not necessarily make it undesirable. Regulating the discriminatory outcomes of algorithms is likely to be welfare-increasing by reducing negative externalities to the populations being discriminated against and overcoming collective action problems. And, like other aspects of antidiscrimination law, it would also have desirable distributive purposes.[313]

This, of course, would not eradicate underlying biases or replace existing proposals of legal reform that address wider problems. Computers will not fix our societal biases, but we can stop them from perpetuating and amplifying them.

CONCLUSION

Algorithmic discrimination continues to puzzle scholars of law and technology. But humans still make the most relevant decisions in algorithmic decision-making processes. Because algorithmic discrimination relies on human decisions, algorithmic and human discrimination work similarly in terms of information availability: depending on the information involved, giving the decision-maker less information can either prevent or worsen both algorithmic and human discrimination.

This Article introduces a preventive approach to algorithmic discrimination. For machine learning algorithms, blocking information can hide and sometimes worsen discrimination. However, while blocking data on protected categories is unhelpful, *shaping* the information that includes protected categories can be effective at eliminating bias from the data that decision-making models are trained with and, in turn, eliminating discrimination from such models. We can edit training data to resemble the more equal world that the law dictates we should live in.

To some extent, the literature on algorithmic discrimination has so far explored legal solutions to an information problem. The method proposed here brings a complementary information-based solution to the information problem. In addition to repairing discrimination after it takes place as traditional antidiscrimination law does, this information-based approach can prevent some instantiations of discrimination.

---

311. *See generally* Jolls, *supra* note 309.

312. Hardt et al., *supra* note 167.

313. *See* John Gardner, *Discrimination as Injustice*, 16 OXF. J. LEG. STUD. 353, 355-56 (1996); John Gardner, *Liberals and Unlawful Discrimination*, 9 OXF. J. LEG. STUD. 1, 11 (1989).

This approach is compatible with, and can be implemented alongside the existing, compensation-based legal solutions. It can prevent disparate impact under a disparate treatment logic, avoiding constitutional challenges; it is a corollary of taking the antisubordination principle seriously in the context of algorithmic discrimination; it can be applied to black box models when other methods cannot; and it has the advantage of preventing discrimination rather than reparing it.

***