Proxy Discrimination in the Age of Artificial Intelligence and Big Data

Anya E.R. Prince & Daniel Schwarcz[*]

ABSTRACT: Big data and Artificial Intelligence ("AI") are revolutionizing the ways in which firms, governments, and employers classify individuals. Surprisingly, however, one of the most important threats to anti-discrimination regimes posed by this revolution is largely unexplored or misunderstood in the extant literature. This is the risk that modern algorithms will result in "proxy discrimination." Proxy discrimination is a particularly pernicious subset of disparate impact. Like all forms of disparate impact, it involves a facially neutral practice that disproportionately harms members of a protected class. But a practice producing a disparate impact only amounts to proxy discrimination when the usefulness to the discriminator of the facially neutral practice derives, at least in part, from the very fact that it produces a disparate impact. Historically, this occurred when a firm intentionally sought to discriminate against members of a protected class by relying on a proxy for class membership, such as zip code. However, proxy discrimination need not be intentional when membership in a protected class is predictive of a discriminator's facially neutral goal, making discrimination "rational." In these cases, firms may unwittingly proxy discriminate, knowing only that a facially neutral practice produces desirable outcomes. This Article argues that AI and big data are game changers when it comes to this risk of unintentional, but "rational," proxy discrimination. AIs armed with big data are inherently structured to engage in proxy discrimination whenever they are deprived of information about membership in a legally suspect class whose predictive power cannot be measured more directly by non-suspect data available to the AI. Simply denying AIs access to the most intuitive proxies for such predictive but suspect characteristics does little to thwart this process; instead it simply causes AIs to locate less intuitive proxies. For these reasons,

[*] Anya E.R. Prince (anya-prince@uiowa.edu) is an Associate Professor, University of Iowa College of Law. Daniel Schwarcz (Schwarcz@umn.edu) is the Fredrikson & Byron Professor of Law, University of Minnesota Law School. For comments and suggestions on preliminary drafts, we thank Ken Abraham, Ronen Avraham, Jessica Clarke, I. Glenn Cohen, James Grimmelman, Jill Hasday, Claire Hill, Dave Jones, Sonia Katyal, Pauline Kim, Kyle Logue, Peter Molk, Chris Odinet, Nicholson Price, Jessica Roberts, Andrew Selbst, Elizabeth Sepper, Rory Van Loo and participants of the Consumer Law Conference at Berkeley Law School.

as AIs become even smarter and big data becomes even bigger, proxy discrimination will represent an increasingly fundamental challenge to anti-discrimination regimes that seek to limit discrimination based on potentially predictive traits. Numerous anti-discrimination regimes do just that, limiting discrimination based on factors like preexisting conditions, genetics, disability, sex, and even race. This Article offers a menu of potential strategies for combatting this risk of proxy discrimination by AIs, including prohibiting the use of non-approved types of discrimination, mandating the collection and disclosure of data about impacted individuals' membership in legally protected classes, and requiring firms to employ statistical models that isolate only the predictive power of non-suspect variables.

I. INTRODUCTION

Big data and Artificial Intelligence ("AI") are revolutionizing the ways in which firms, governments, and employers classify individuals.[1] Insurers, for instance, increasingly set premiums based on complex algorithms that process massive amounts of data to predict future claims.[2] Prospective employers deploy AI and big data to decide which applicants to interview or hire.[3] And various actors within the criminal justice system—ranging from police departments to judges—now use predictive analytics to guide their decision-making.[4]

---

1. We use the term "artificial intelligence" to encompass a broad array of computational techniques for predicting future outcomes based on analysis of past data. These techniques include "machine learning," "deep learning," "learning algorithms," and many other terms. While there are often important differences among these various types of AIs, these distinctions are not pertinent to the analysis in this Article.

2. *See* Rick Swedloff, *Risk Classification's Big Data (R)evolution*, 21 CONN. INS. L.J. 339, 340–44 (2014); Herb Weisbaum, *Data Mining Is Now Used to Set Insurance Rates; Critics Cry Foul*, CNBC (Apr. 16, 2014, 11:29 AM), https://www.cnbc.com/2014/04/16/data-mining-is-now-used-to-set-insurance-rates-critics-cry-fowl.html [https://perma.cc/MQ28-C8RA]; *see also* Ray Lehmann, *Why 'Big Data' Will Force Insurance Companies to Think Hard About Race*, INS. J. (Mar. 27, 2018), https://www.insurancejournal.com/blogs/right-street/2018/03/27/484530.htm [https://perma.cc/4GBZ-MBZZ] ("According to a 2015 survey conducted by Willis Towers Watson, 42 percent of executives from the property and casualty insurance industry said they were already using big data in pricing, underwriting and risk selection, and 77 percent said they expected to do so within two years.").

3. *See* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 860 (2017) ("Employers are increasingly relying on data analytic tools to make personnel decisions . . . .").

4. *See* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1068–76 (2019); Elizabeth E. Joh, *Policing by Numbers: Big Data and the Fourth Amendment*, 89 WASH. L. REV. 35, 42–55 (2014); Sharad Goel, Ravi Shroff, Jennifer Skeem & Christopher Slobogin, The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment 1 (Dec. 26, 2018)

This big data revolution raises numerous complex challenges for anti-discrimination regimes.[5] Perhaps most obviously, improperly-designed algorithms or errant data can disproportionately harm discrete subsets of the population.[6] But even correctly programmed algorithms armed with accurate data can reinforce past discriminatory patterns.[7] Surprisingly, however, one of the most important threats to anti-discrimination regimes posed by big data and AI is largely unexplored or misunderstood in the extant legal literature. This is the risk that modern AIs will result in "proxy discrimination."

Proxy discrimination is a particularly pernicious subset of disparate impact. Like all forms of disparate impact, it involves a facially neutral practice that disproportionately harms members of a protected class.[8] But a practice producing a disparate impact only amounts to proxy discrimination when a

(unpublished manuscript), *available at* https://ssrn.com/abstract=3306723 [https://perma.cc/4DFC-2K6U]. Of course, these examples hardly exhaust the scope and import of AI and Big Data. For instance, these forces are fundamentally reshaping the consumer credit economy. *See* Matthew Adam Bruckner, *The Promise and Perils of Algorithmic Lenders' Use of Big Data*, 93 CHI.-KENT L. REV. 3, 11–15 (2018); Christopher K. Odinet, *Consumer Bitcredit and Fintech Lending*, 69 ALA. L. REV. 781, 802–04 (2018). They are also fundamentally changing the business of financial advice, offering personalized AI assistants that promise to improve consumer decision-making. *See* Rory Van Loo, *Digital Market Perfection*, 117 MICH. L. REV. 815, 862–63, 878–79 (2019).

5. *See generally* CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016) (discussing how algorithms used in society can perpetuate discrimination, in part through perpetuation of disadvantage); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 682 (2016) [hereinafter Barocas & Selbst, *Big Data*] (discussing how data is often imperfect and therefore algorithms inherit the prejudice of the original decision makers); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 99–101 (2014) (discussing ways that predictive analytic tools can perpetuate discriminatory practices).

6. *See* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4–5 (2014) (describing how human beings programming automated systems can lead to inaccurate results because the source code, predictive algorithms and datasets may contain human biases that have a disparate impact on certain groups).

7. *See* Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 524–28 (2018) (arguing that "facially neutral" algorithms producing unequal outcomes should be challenged as violating Title VII's stereotype theory of liability).

8. None of this is to suggest that mere disparate impact alone is not a significant issue raised by big data and algorithms. *See, e.g.*, Robert Bartlett, Adair Morse, Richard Stanton & Nancy Wallace, *Consumer-Lending Discrimination in the FinTech Era* 4 (Nat'l Bureau of Econ. Research, Working Paper No. 25943, 2019) (finding that disparate impact extracts as much rents as face-to-face discrimination). But the issue of whether disparate impact alone should be actionable is distinct from the issue of proxy discrimination. For arguments about the desirability of disparate impact in insurance, see generally Matthew Jordan Cochran, *Fairness in Disparity: Challenging the Application of Disparate Impact Theory in Fair Housing Claims Against Insurers*, 21 GEO. MASON U. C.R. L.J. 159 (2011) (discussing the use of disparate impact theory under Title VII and potential applicability to Fair Housing Act claims against insurers); Dana L. Kaersvang, Note, *The Fair Housing Act and Disparate Impact in Homeowners Insurance*, 104 MICH. L. REV. 1993 (2006) (providing additional analysis on the application of the Fair Housing Act's disparate impact standard to insurance); and Ronen Avraham, A Normative Theory for Insurance Antidiscrimination Law (Jan. 2019) (unpublished manuscript) (on file with Author) (offering a framework for evaluating the costs and benefits of insurance discrimination laws).

second condition is met. In particular, proxy discrimination requires that the usefulness to the discriminator of a facially neutral practice derives, at least in part, from the very fact that it produces a disparate impact.[9] This condition can be met either when the discriminator intends to disparately impact a protected group or when a legally-prohibited characteristic is predictive of the discriminator's goals in ways that cannot be captured more directly by non-suspect data.

This distinction between generalized disparate impact and the more specific phenomenon of proxy discrimination is well illustrated by positing a life insurer that uses an AI to price its policies. Suppose that the model generated by the insurer's AI charges more for coverage to applicants who are members of a Facebook group focused on increasing the availability to African Americans' of genetic testing for *BRCA* variants, which are highly predictive of certain cancers.[10] In these circumstances, the insurer would almost certainly be proxy discriminating for genetic information. First, the AI's pricing model would presumably disparately impact those with a genetic predisposition to breast and ovarian cancer, as members of the Facebook group are relatively likely to have a family connection to these *BRCA*-related cancers. Second, this link between membership in the Facebook group and genetic history would hardly be fortuitous. To the contrary, it would presumably be the very reason why the AI latched on to membership in the Facebook group when setting applicants' premiums.[11] Framing the point in

9. *See generally* James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164 (2017) (presenting a hypothetical as a teaching tool to showcase disparate impact and proxies); Darcy Steeg Morris, Daniel Schwarcz & Joshua C. Teitelbaum, *Do Credit-Based Insurance Scores Proxy for Income in Predicting Auto Claim Risk?*, 14 J. EMPIRICAL LEGAL STUD. 397, 418–21 (2017) (showing that one insurer's use of credit-based insurance scores does not have a disparate impact based on income and therefore does not operate as a proxy for income); *see generally also* Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J. 206, 209 (2011) (discussing how the FTC examined credit score use in auto-insurance pricing as a proxy for race).

10. In fact, according to a complaint recently lodged with the FTC, a vulnerability in Facebook private groups means that information about who is in what private group could be scraped by an algorithm. For the text of the report and Facebook's reply, see *Facebook Patient FTC Complaints: Released 2/18/19*, MISSING FACEBOOK PATIENT CONSENT, https://missingconsent.org/facebook-patient-ftc-complaints [https://perma.cc/EC85-PMLS].

11. State law is inconsistent regarding the rules that govern the use of genetic information by life insurers, disability insurers, and long-term care insurers. By contrast, the federal Genetic Information and Non-Discrimination Act ("GINA") prohibits health insurers and employers from discriminating on the basis of such genetic information. Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233, 122 Stat. 881 (codified as amended in scattered sections of 26, 29, and 42 U.S.C.); *see* Robert Klitzman, Paul S. Appelbaum & Wendy Chung, *Should Life Insurers Have Access to Genetic Test Results?*, 312 JAMA 1855, 1855–56 (2014) (arguing that modest life insurance coverage should be available without underwriting based on genetic information). The rise in genetic testing has created many legal questions. *See generally* Susan M. Wolf & Jeffrey P. Kahn, *Genetic Testing and the Future of Disability Insurance: Ethics, Law & Policy*, 35 J.L. MED. & ETHICS 6 (2007) (describing the role and problems of using genetic testing in disability insurance); Leslie E. Wolf, Erin Fuse Brown, Ryan Kerr, Genevieve Razick, Gregory Tanner, Brett

econometric terms, data on applicants' membership in the Facebook group would likely cease to be predictive of claims in a model that controlled for applicants' genetic predispositions to cancer.[12]

By contrast, the insurer in this example would likely not be proxy discriminating with respect to policyholder race, even if African Americans were disproportionately harmed by the insurer's actions. To be sure, it is plausible to assume that the insurer's actions disparately impacted African Americans given the race-specific nature of the Facebook group. Even so, the predictive power of applicants' membership in the group would probably have nothing to do with the correlation between such membership and applicants' race. Instead, the disparate impact felt by African Americans would be merely fortuitous. Once again framing this point in econometric terms, applicants' membership in the Facebook group would be equally predictive of future insurance claims even in a model that controlled for applicants' race, assuming that any differences in life expectancy between African-Americans and other applicants can be explained by variables like income or access to healthcare.

Historically, proxy discrimination was generally understood as a type of intentional discrimination, rather than as a subset of disparate impact. Indeed, the paradigmatic example of proxy discrimination by humans involves financial firms that refused to serve predominantly African American geographic regions, a phenomenon known as redlining. This practice constituted intentional proxy discrimination because the disparate impact it produced was by design: The usefulness to firms of refusing to serve redlined geographic regions was that it allowed them to covertly achieve their discriminatory aims.

However, proxy discrimination need not be intentional when membership in a protected class is predictive of a discriminator's legitimate goal, making discrimination "rational."[13] In these cases, firms may unwittingly proxy discriminate, knowing only that a facially-neutral practice produces desirable outcomes. The insurance example above is once again illustrative. The insurer in this example presumably programmed its AI simply to minimize future claims. It might be unaware that the AI was targeting applicants' membership in a Facebook group to achieve this objective. And even if the insurer was so aware, it likely would not know that Facebook groups were predictive of genetic risk because they indirectly captured genetic

---

Duvall, Sakinah Jones, Jack Brackney & Tatiana Posada, *The Web of Legal Protections for Participants in Genomic Research*, 29 HEALTH MATRIX 1 (2019) (examining the various state and federal legal protections provided to participants in genomic research, including in life, long-term care, and disability insurance).

12. *See generally* Pope & Sydnor, *supra* note 9 (discussing how proxy effects could be eliminated utilizing statistical methods).

13. Mark A. Rothstein & Mary R. Anderlik, *What Is Genetic Discrimination, and When and How Can It Be Prevented?*, 3 GENETICS MED. 354, 354–55 (2001).

information.[14] Either way, the insurer would be engaging in unintentional proxy discrimination, at least assuming—as we do throughout this Article—that an AI cannot intentionally discriminate independently of any human.[15]

Unintentional proxy discrimination by human actors is uncommon and can typically be prevented by scrutinizing use of obvious potential proxies for membership in a protected group, like zip code.[16] But unintentional proxy discrimination by AIs is virtually inevitable whenever the law seeks to prohibit discrimination on the basis of traits containing predictive information that cannot be captured more directly within the model by non-suspect data; a type of information we label as "directly predictive."[17] The inherent tendency of AIs to engage in proxy discrimination when they are deprived of directly predictive traits follows inextricably from their structure.[18] Predictive AIs are programmed to locate correlations between input data and target variables of interest. But unlike traditional statistical models, AIs do not accomplish this by relying on a human's starting intuition about causal explanations for

14. For a useful breakdown of the different types of opacity implicated by machine learning algorithms, see Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1089–99 (2018) [hereinafter Selbst & Barocas, *Intuitive Appeal*]. In Selbst and Barocas's terms, there are several reasons for the insurer's ignorance. First, the algorithm requires specialized knowledge to understand. *Id.* In some cases, the insurer's employees may not have sufficient expertise to comprehend how or why the algorithm is producing prices for different customers.

Second, the model that the algorithm produces may be so complex and sophisticated that it is "inscrutable" even for those within the company that possess the necessary expertise. *Id.* In other words, the sheer complexity of the algorithm may prevent those within the insurer from understanding how the model operates. As applied here, the model's complexity may prevent the insurer from seeing the link between visits to the specified website and the higher rates produced by the AI's model.

Finally, even if the insurer understands how the model operates, it may not understand why it operates the way it does. This is a scenario that Selbst and Barocas label as the "nonintuitive" nature of algorithms. *Id.* at 1091. As applied here, the insurer may indeed know that its AI suggests higher prices for those who visit the website at issue, but not realize that the explanation for this fact derives from the website's capacity to proxy for genetic information. *See generally* Matt Turek, *Explainable Artificial Intelligence (XAI)*, DEF. ADVANCED RES. PROJECTS AGENCY, www.darpa.mil/program/explainable-artificial-intelligence [https://perma.cc/2DS4-ZMNB] (discussing the interpretability of algorithms); *Making Computers Explain Themselves*, MIT COMPUT. SCI. & ARTIFICIAL INTELLIGENCE LAB (Oct. 27, 2016), www.csail.mit.edu/making_computers_explain_themselves [https://perma.cc/C97B-PGUR] (explaining the importance of understanding algorithm decision-making).

15. As a doctrinal matter, this seems likely. *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 699 (discussing how discriminatory data mining is analogous to unintentional disparate impact analysis); Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395, 404 (2018). *But see* Bornstein, *supra* note 7, at 535 (arguing that algorithmic discrimination at large could fall under the anti-stereotyping concept within Title VII's disparate treatment).

16. For a discussion of this point in the insurance context, see Daniel Schwarcz, *Ending Public Utility Style Rate Regulation in Insurance*, 35 YALE J. ON REG. 941, 978 (2018).

17. *See infra* Section II.C.1.

18. *See* Kim, *supra* note 3, at 898–99.

statistical linkages between input data and the target variable.[19] Instead, AIs use training data to discover on their own what characteristics can be used to predict the target variable.[20] Although this process completely ignores causation, it results in AIs inevitably "seeking out" proxies for directly predictive characteristics when data on these characteristics is not made available to the AI due to legal prohibitions.[21] Simply denying AIs access to the most intuitive proxies for directly predictive variables does little to thwart this process; instead it simply causes AIs to produce models that rely on less intuitive proxies.

Thus, this Article's central argument is that as AIs become even smarter and big data becomes even bigger, proxy discrimination will represent an increasingly fundamental challenge to anti-discrimination regimes[22] that seek to prohibit discrimination based on directly predictive traits.[23] Such prohibitions on the use of directly predictive characteristics are particularly important in insurance regulation.[24] For instance, the Patient Protection and Affordable Care Act ("ACA") prohibits insurers from discriminating on the basis of health status[25] and the Genetic Information Nondiscrimination Act ("GINA") prohibits discrimination by covered health insurers (and employers, who often provide health insurance) on the basis of genetic information.[26] However, legally-suspect characteristics are directly predictive of seemingly neutral goals outside of the insurance setting as well. Thus, employers are prohibited from considering sex, race, age, and disability in

19. *See* Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 19–20 (2002) (explaining the traditional method of empirical research).

20. *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 677–78; Kim, *supra* note 3, at 878–80; Machine learning algorithms generate their own models to predict future outcomes based on analysis of training data. *See* Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter & Luciano Floridi, *The Ethics of Algorithms: Mapping the Debate*, BIG DATA & SOC'Y, July–Dec. 2016, at 3. For more on how machine-learning algorithms operate, see *infra* Section II.B.

21. *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 691–92.

22. *See* Lior Jacob Strahilevitz, *Privacy Versus Antidiscrimination*, 75 U. CHI. L. REV. 363, 365–67 (2008) (discussing "rational racism").

23. Directly predictive data is predictive of a target variable (i.e., minimized future predicted claims) with training data that is both correctly "labelled" and "collected." *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 677–78.

24. *See* Kenneth S. Abraham, *Efficiency and Fairness in Insurance Risk Classification*, 71 VA. L. REV. 403, 407–08 (1985) [hereinafter Abraham, *Efficiency and Fairness*] (describing the adverse consequences of insurance competition, pricing and risk classification).

25. Individual and Group Market Reforms, 42 U.S.C. §§ 300gg–300gg-2 (2012); *see* JESSICA L. ROBERTS & ELIZABETH WEEKS, HEALTHISM: HEALTH-STATUS DISCRIMINATION AND THE LAW 112–13 (2018).

26. Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233, 122 Stat. 881 (codified as amended in scattered sections of 26, 29, and 42 U.S.C); *see* Anya E.R. Prince, *Insurance Risk Classification in an Era of Genomics: Is a Rational Discrimination Policy Rational?*, 96 NEB. L. REV. 624, 626 (2018); *see also* Bradley A. Areheart & Jessica L. Roberts, *GINA, Big Data, and the Future of Employee Privacy*, 128 YALE L.J. 710, 716 (2019) (explaining the basics of the GINA law).

hiring decisions, even though these factors can be directly predictive of neutral objectives, like maximizing employee hours worked or total sales.[27]

Proxy discrimination by AIs is most likely to occur when prohibited traits are directly predictive of legitimate outcomes in ways that cannot be more directly captured by alternative data. For that reason, proxy discrimination by AIs may, at first blush, seem normatively acceptable. It is not. This is because laws that seek to prohibit discrimination on the basis of directly predictive traits—the only types of laws that inevitably tend to produce proxy discrimination by AIs—are motivated principally by the goal of preventing specific outcomes for members of protected groups. Unlike many other types of anti-discrimination laws, the questions of how or why bad outcomes obtain for these groups are generally secondary; that is precisely why these laws prohibit discrimination even when it is rational, rather than only when it is a byproduct of animus or irrelevant stereotypes. Proxy discrimination by AIs strikes at the heart of this outcome-oriented goal. To illustrate, such discrimination could result in individuals who get troubling genetic test results finding it harder to secure employment or in women who report being victimized by domestic abuse finding it more difficult to purchase life insurance. These results are normatively troubling irrespective of how or why they come to fruition.

Despite the substantial risks associated with proxy discrimination by AIs, most of the extant legal literature and public policy analysis on AI fails to clearly distinguish between proxy discrimination and ordinary disparate impact analysis.[28] Instead, most analyses conflate scenarios in which an

27. *See* Grutter v. Bollinger, 539 U.S. 306, 326 (2003) (holding that a law school's race-conscious admissions policy violated challengers' equal protection rights).

28. The clearest exception is the excellent piece *Incomprehensible Discrimination* by Grimmelmann & Westreich. Grimmelmann & Westreich, *supra* note 9. This unique piece is styled as a vignette and mock judicial opinion, focusing on analyzing a specific fictional case (an employment-based AI employed in the fictional universe of the movie *Zootopia*) under a specific legal regime (Title VII). It ultimately resolves this fictional case by making the same distinction between proxy discrimination and disparate impact we focus on in this Article. *See id.* at 170 ("The problem is that there is no explanation in the record as to which of these two correlations, if either, is causal. It may be that the factors directly measure applicant characteristics that determine success in the challenging and dangerous field of police work, and that those characteristics happen to be unequally distributed in our diverse society. It may also be that these factors are instead measuring applicants' species and that they measure likely job performance *only because they are identifying species* in an applicant pool where the relevant characteristics are unequally distributed."). For this reason, the piece does not attempt to systematically explore the unique dangers of proxy discrimination by AIs or how those dangers might play out and be addressed across different anti-discrimination regimes. At least one other article briefly refers to the possibility that an AI might engage in proxy discrimination, without systematically analyzing this possibility. *See* Sullivan, *supra* note 15, at 406–07 ("That is, suppose that, to avoid this problem, Arti is programmed not to use protected traits in its operations. While it would then be race- and gender-blind, faithfulness to its mission would seem to require it to look to 'neutral' criteria but ones with a high correlation to the now-off-limits prohibited characteristics."). Several prior works clearly explain the distinction between disparate impact and proxy discrimination by

algorithm latches on to a variable that fortuitously happens to be correlated with membership in a suspect class, and scenarios in which an algorithm uses a variable whose predictive power derives from its correlation with membership in the suspect class.[29] This Article clarifies that only the latter is proxy discrimination, suggests that this phenomenon is particularly pernicious, and argues that the continued evolution of AI and big data will cause proxy discrimination to increase substantially whenever anti-discrimination law seeks to prohibit the use of characteristics that are directly predictive of risk.

For these reasons, anti-discrimination laws that prohibit discrimination based on directly predictive characteristics must adapt to combat proxy discrimination in the age of AI and big data. This Article offers a menu of potential strategies for achieving this objective. For instance, impacted anti-discrimination regimes could allow, and perhaps even require, that firms using predictive AIs collect data about individuals' potential membership in legally protected classes. In some cases, this data should be shared with regulators and/or disclosed to the public in summary form.[30] Such data is necessary for firms, regulators, litigants, and others to test whether any particular AI is, in fact, engaging in proxy discrimination.[31] Alternatively, anti-discrimination regimes could develop specific criteria for requiring firms that

---

algorithm, though they do not consider it in the context of AI and do not focus substantial attention on the distinction. *See infra* note 44 and accompanying text. *See generally, e.g.*, Pope & Sydnor, *supra* note 9 (describing the circumstances in which proxy discrimination occurs); Steeg Morris, Schwarcz & Teitelbaum, *supra* note 9, at 420 (describing disparate impact).

29. *See, e.g.*, Barocas & Selbst, *Big Data*, *supra* note 5, at 691 ("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. In other words, the very same criteria that correctly sort individuals according to their predicted likelihood of excelling at a job—as formalized in some fashion—may also sort individuals according to class membership."); Kim, *supra* note 3, at 877 ("Data models may also discriminate when neutral factors act as 'proxies' for sensitive characteristics like race or sex. Those neutral factors may be highly correlated with membership in a protected class, and also correlate with outcomes of interest. In such a situation, those neutral factors may produce results that systematically disadvantage protected groups, even though the model's creators have no discriminatory intent, and the sensitive characteristics have been removed from the data."). *See also generally* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019) (explaining that restricting data models use of sensitive characteristics simply leads to the models using neutral factors to discriminate); Cass R. Sunstein, *Algorithms, Correcting Biases*, SOC. RES. (forthcoming) (analyzing whether discrimination by algorithms produces disparate treatment or disparate impact without identifying unique issues associated with proxy discrimination).

30. *See* Prohibit Auto Insurance Discrimination Act, H.R. 5502, 115th Cong. (2d Sess. 2018) (prohibiting auto insurers from taking education, occupation, employment, homeownership, credit score, and various other information into consideration when determining insurance rates or eligibility).

31. *See* Kim, *supra* note 3, at 898, 916–18 (discussing data classification bias and the use of proxies).

are at substantial risk of engaging in proxy discrimination to deploy "ethical algorithms" that explicitly seek to eliminate the capacity of any facially-neutral considerations to proxy for prohibited characteristics.[32] Yet a third option for combatting proxy discrimination would be to flip the default approach to anti-discrimination law, such that all forms of discrimination are prohibited except those that are specifically allowed.[33] Approved forms of discrimination could then be set by statute or regulation based on evidence regarding the risk of proxy discrimination.

In advancing these arguments, this Article proceeds as follows. Part II begins by tracing the evolution of proxy discrimination from a form of shrouded intentional discrimination by human actors to its modern and future incarnation in AIs. It explains why proxy discrimination by AIs is inevitable when the law seeks to prohibit discrimination based on directly predictive traits, and when anti-discrimination rules meet this initial condition. Having laid these foundations in Part II, Part III identifies the anti-discrimination regimes that are most at risk of proxy discrimination by AIs because they target characteristics that are directly predictive of discriminators' otherwise valid objectives. Part III also explains why proxy discrimination by AIs in these settings is so normatively troubling. Finally, Part IV highlights how current law is inadequate to address proxy discrimination by AIs and explores potential responses to this risk, drawing from several nascent efforts to shield existing anti-discrimination regimes from the unique risks associated with the growth of AI.

II. PROXY DISCRIMINATION BY HUMANS AND AIS

Proxy discrimination occurs when a facially-neutral trait is utilized as a stand-in—or proxy—for a prohibited trait. Historically, firms engaged in proxy discrimination in an intentional effort to thwart anti-discrimination laws. However, proxy discrimination need not be intentional when the law prohibits "rational" or "statistical" discrimination, where discrimination can be justified by genuine statistical differences in relevant expected outcomes among members of different groups. When this initial condition is met, firms may unintentionally discriminate on the basis of facially-neutral proxies for protected traits simply because doing so "works" to help the firm achieve legitimate objectives. Section II.A of this Part explains these points in more detail.

---

32. *See* Pope & Sydnor, *supra* note 9, at 207–09. For practical proposals to implement ethical algorithms in insurance, see Birny Birnbaum, Exec. Dir., Ctr. for Econ. Justice, Presentation at CAS Ratemaking Seminar: Insurance Regulation: The Challenge of Big Data in Insurance (March 20, 2018) (on file with Author).

33. *See, e.g.*, Harvey Rosenfield, *Auto Insurance: Crisis and Reform*, 29 U. MEM. L. REV. 69, 129 (1998) ("Notwithstanding any other provision of law, the use of any criterion without such approval shall constitute unfair discrimination.").

Section II.B then explores how the emergence of decision-making by AIs will dramatically alter the character of proxy discrimination when the law seeks to prohibit a specific type of rational discrimination. In particular, whenever the law seeks to prohibit discrimination based on traits whose predictive power cannot be measured more directly by facially-neutral data that is available to the AI ("directly predictive" data), then AIs will inevitably engage in increasingly effective proxy discrimination.

Finally, Section II.C explores in more detail when illicit traits will be directly predictive of legitimate outcomes, thus creating the likelihood of proxy discrimination by AIs. This risk is greatest when a legally-prohibited trait is causally linked to a desired outcome, as is the case with genetic information and preexisting conditions. But it is also substantial when a suspect trait is directly linked to desired outcomes for reasons that are opaque, such that the trait's predictive power is not mediated through presently quantifiable or available information. By contrast, the risk of proxy discrimination by AIs is lowest when legally suspect traits are only "indirectly predictive" of legitimate outcomes, meaning that they proxy for another quantifiable and potentially available variable, like college graduation. By decreasing the cost of acquiring and processing individualized data that most directly matters to outcomes, AIs can actually limit the risk of the latter, indirect "rational stereotyping."[34]

A. PROXY DISCRIMINATION BY HUMAN ACTORS

Proxy discrimination is not a new phenomenon.[35] Historically, the term referred to deliberate attempts to indirectly discriminate against protected groups. This type of intentional proxy discrimination occurs whenever an actor discriminates based on a facially-neutral characteristic that is correlated with membership in a legally protected group and that discrimination is motivated by the discriminator's knowledge of this correlation.[36] The tighter this correlation, the more effectively the discriminator can achieve its ultimate goal of weeding out members of the targeted protected group. Meanwhile, because the discriminator never explicitly considers membership in a protected group as part of its decision-making process, it can claim that it is complying with applicable anti-discrimination rules.

The classic example of intentional proxy discrimination is redlining by financial institutions.[37] During the mid-Twentieth Century, various state and

34. *See infra* Section IV.B.2.

35. *See, e.g.*, Larry Alexander & Kevin Cole, *Discrimination by Proxy*, 14 CONST. COMMENT. 453, 453 (1997) (analyzing the use of proxies under the anti-discrimination, disparate impact, and intent principles of constitutional law); Deborah Hellman, *Two Types of Discrimination: The Familiar and the Forgotten*, 86 CALIF. L. REV. 315, 317–18 (1998).

36. *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 691–92, 694.

37. *See* MEHRSA BARADARAN, THE COLOR OF MONEY: BLACK BANKS AND THE RACIAL WEALTH GAP 105–06 (2017). *See generally* Gregory D. Squires, *Racial Profiling, Insurance Style: Insurance Redlining and the Uneven Development of Metropolitan Areas*, 25 J. URB. AFF. 391 (2003) (examining

federal laws were passed prohibiting financial institutions like banks and insurers from discriminating on the basis of race.[38] Rather than continue to explicitly consider race in their underwriting and pricing decisions, many financial institutions resorted to proxy discrimination by refusing to serve geographic areas that were predominantly African American.[39] Although financial institutions publicly claimed that such redlining was motivated by concerns having nothing to do with race, in many cases quite the opposite was true: These firms specifically sought to limit their African American customers by discriminating on the basis of an obvious proxy for race.[40]

Intentional proxy discrimination clearly violates most anti-discrimination laws because it constitutes disparate treatment. Disparate treatment occurs

---

the role of racial profiling in the property insurance industry and its contribution to racial segregation). The historical link between proxy discrimination and discriminatory intent is also nicely illustrated by the Supreme Court case *Hazen Paper Co. v. Biggins*. Hazen Paper Co. v. Biggins, 507 U.S. 604 (1993). The issue in *Biggins* was whether an employee who had been fired because his pension was close to vesting could successfully advance a disparate treatment claim under the Age Discrimination in Employment Act ("ADEA"). *See id.* at 608; Age Discrimination in Employment Act of 1967, Pub. L. No. 90-202, 81 Stat. 602 (codified as amended at 29 U.S.C. §§ 621–634 (2012)). In concluding that he could not, the Court emphasized that years of service (a non-suspect classifier under the ADEA) was analytically distinct from age (a prohibited characteristic under ADEA), notwithstanding the fact that the two were obviously correlated with one another. *Biggins*, 507 U.S. at 612. At the same time, the court clarified that the case would be different if there were evidence that the "employer . . . target[ed] employees with a particular pension status on the assumption that these employees are likely to be older." *Id.* In that event, "[p]ension status may be a proxy for age . . . in the sense that the employer may suppose a correlation between the two factors and act accordingly." *Id.* at 613.

38. *See* MICHAEL S. BARR, HOWELL E. JACKSON & MARGARET E. TAHYAR, FINANCIAL REGULATION: LAW AND POLICY 60 (2d ed. 2016) (reviewing laws prohibiting discrimination on the basis of race in credit). *See generally* Ronen Avraham, Kyle D. Logue & Daniel Schwarcz, *Understanding Insurance Antidiscrimination Laws*, 87 S. CAL. L. REV. 195 (2014) (reviewing prohibitions against insurers' consideration of race in insurance).

39. *See, e.g.*, Squires, *supra* note 37, at 396–97.

40. For instance, insurance textbooks from the 1950s warned underwriters of the importance of determining applicants' race and ethnicity in assessing their riskiness. Brian J. Glenn, *Post-Modernism: The Basis of Insurance*, 6 RISK MGM'T & INS. REV. 131, 134 (2003). As one commentator explained in the late 1970s:

> Although the core concern of the underwriter is the human characteristics of the risk, cheap screening indicators are adopted as surrogates for solid information about the attitudes and values of the prospective insured. . . . Even generalized underwriting texts include occupational, ethnic, racial, geographic, and cultural characterizations certain to give offense if publicly stated.

Robert Works, *Whatever's FAIR—Adequacy, Equity, and the Underwriting Prerogative in Property Insurance Markets*, 56 NEB. L. REV. 445, 471 (1977) (citation omitted); *see also* Regina Austin, *The Insurance Classification Controversy*, 131 U. PA. L. REV. 517, 537–38 (1983) (describing insurers' reliance on occupational and cultural stereotypes without any empirical support for these stereotypes). Studies show that such redlining did not, in fact, accurately reflect the riskiness of the affected areas. *See generally* Robert W. Klein, *Availability and Affordability Problems in Urban Homeowners Insurance Markets*, *in* INSURANCE REDLINING: DISINVESTMENT, REINVESTMENT, AND THE EVOLVING ROLE OF FINANCIAL INSTITUTIONS (Gregory D. Squires ed., 1997) (examining the lack of statistical support underlying the use of redlining).

when a discriminator intentionally treats an individual less favorably than others because of a protected trait.[41] Although such disparate treatment is most closely associated with employment anti-discrimination laws, it constitutes a paradigmatic violation of virtually all anti-discrimination regimes —including the laws governing employment, insurance, housing, and banking. When a firm intentionally discriminates on the basis of a characteristic because it is a proxy for a protected characteristic, it undoubtedly targets members of a protected group for less favorable treatment and violates these laws.

Despite the historical link between proxy discrimination and discriminatory intent, proxy discrimination need not be intentional. Instead, humans can unwittingly proxy discriminate when the law prohibits "rational discrimination" that can be justified based on statistical differences among protected and unprotected groups.[42] In these circumstances, a person or firm may find that discrimination based on a facially-neutral characteristic is predictive of its legitimate objectives, even though the characteristic's predictive power derives from its correlation with a legally-prohibited characteristic.[43] This would constitute proxy discrimination, because it would (1) disparately impact members of a protected group, and (2) prove useful to the firm for precisely this reason. Yet the unwitting discriminator may be unaware of these realities, realizing only that discrimination based on a facially neutral practice "works" to predict a legitimate goal, like minimizing future insurance claims.

---

41. *See* Ricci v. DeStefano, 557 U.S. 557, 577 (2009) (holding that disparate treatment occurs when "an employer has 'treated [a] particular person less favorably than others because of" a protected trait." (quoting Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 987 (1988))).

42. There is a vast legal and economic literature on rational discrimination. For some strong illustrative examples, see, e.g., David Charny & G. Mitu Gulati, *Efficiency-Wages, Tournaments, and Discrimination: A Theory of Employment Discrimination Law for "High-Level" Jobs*, 33 HARV. C.R.-C.L. L. REV. 57, 64–66, 78–85 (1998). *See generally, e.g.*, Amanda Agan & Sonja Starr, *Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment*, 133 Q.J. ECON. 191 (2018) (arguing that "Ban the Box" policies which restrict employers from asking about applicants' criminal backgrounds encourage racial discrimination); Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659 (1972) (discussing the statistical theory of racial discrimination which lead misinformed employers to discriminate against minorities in an effort to maximize profits); Rothstein & Anderlik, *supra* note 13, at 354–55 (discussing the origins and definitions surrounding genetic discrimination).

43. Statistical proxy discrimination phenomenon has received extended treatment in at least two economics articles, in part because it can be easily framed in econometric terms. *See* Pope & Sydnor, *supra* note 9; Steeg Morris, Schwarcz & Teitelbaum, *supra* note 9, at 420. As Pope & Sydnor explain: "Econometrically the problem here is simply classic omitted variable bias. If a variable (e.g., zip code) in the model is correlated with a predictive characteristic that is left out of the model (e.g., race), the included variable will partially proxy for the omitted characteristic and the estimated impact of the included variable will be biased." Pope & Sydnor, *supra* note 9, at 207.

To illustrate, consider how insurers' use of credit information to price coverage could amount to unintentional proxy discrimination.[44] Auto and homeowners insurers routinely set premiums using credit information, which is predictive of future claims. Critics often allege that this practice amounts to proxy discrimination for race and income.[45] This criticism is facially plausible (if not empirically supported) for two reasons.[46] First, insurers' use of credit information almost certainly disparately impacts low-income and minority policyholders, who disproportionately have relatively low credit scores. Second, the reason why credit information predicts future insurance claims could plausibly stem from its capacity to proxy for policyholder income or race, even though insurer discrimination on these bases is generally prohibited. Policyholder income, in particular, might be predictive of future insurance claims if low-income policyholders are more likely to file claims even when losses are only moderately above their deductible.

Notwithstanding the possibility that insurers' use of credit information to price coverage might amount to proxy discrimination, insurers are almost certainly not *intentionally* proxy discriminating against low income or minority policyholders. From insurers' perspectives, incorporating credit information into their statistical models helps predict the legitimate metric of future insurance claims. Some insurers might not even know there is a correlation between the proxy variable (credit scores) and the suspect variable (race and income). And even if insurers are aware of this correlation, they may not believe that this correlation helps to explain the power of credit information to predict claims. Instead, they may believe, as much available evidence in fact indicates, that credit information is predictive of claims because it measures policyholder care levels.[47]

As this example suggests, the distinction between intentional and unintentional proxy discrimination ultimately turns on why a disparate impact produced by a facially neutral practice proves useful to the discriminator. A firm engaging in intentional proxy discrimination finds the

---

44. For an overview of state rules regarding discrimination based on income in insurance, see FED. TRADE COMM'N, CREDIT-BASED INSURANCE SCORES: IMPACTS ON CONSUMERS OF AUTOMOBILE INSURANCE 17–20 (2007), *available at* https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf [https://perma.cc/RXR4-G58T] [hereinafter FTC REPORT].

45. *See* Press Release, Representative Rashida Tlaib, Congresswomen Take Steps to Prevent Automotive Insurance Discrimination with the PAID Act (July 12, 2019), *available at* https://tlaib.house.gov/media/press-releases/congresswomen-take-steps-prevent-automotive-insurance-discrimination-paid-act [https://perma.cc/6AFU-EAWM].

46. *See* Steeg Morris, Schwarcz & Teitelbaum, *supra* note 9, at 403 (describing how insurance scores could plausibly act as a proxy for race and income); FTC REPORT, *supra* note 44, at 61.

47. *See* Steeg Morris, Schwarcz & Teitelbaum, *supra* note 9, at 403 ("[M]any have offered explanations, most often arguing that people with poor credit scores are less careful or responsible in general"); FTC REPORT, *supra* note 44, at 31.

disparate impact produced by its facially neutral practice useful for the simple reason that it helps the firm to stealthily achieve its discriminatory aim.[48] By contrast, the disparate impact produced by unintentional proxy discrimination is useful because it helps a firm achieve a legitimate objective, like predicting future insurance claims.

Unlike intentional proxy discrimination, unintentional proxy discrimination is typically analyzed under a disparate impact framework because the lack of discriminatory intent undermines a disparate treatment claim. However, the availability of a disparate impact theory varies substantially by anti-discrimination regime; while such liability is recognized in the federal regimes governing employment and housing, for instance, it is not generally available under state insurance laws.[49] Where it is available, disparate impact does not require any showing of discriminatory intent, even though such intent may in fact be present.[50] Instead, it requires simply that a facially-neutral practice disproportionately impacts members of a protected group.[51] If so, then the burden shifts to the discriminator to demonstrate that its practice has a legitimate non-discriminatory purpose that is rooted in business necessity. Even if the firm or actor can meet this burden, it may still be in violation of the law if it could achieve its legitimate aims with a less discriminatory alternative.

Figure 1, below, visually lays out the relationship among intentional proxy discrimination, unintentional proxy discrimination, disparate impact, and disparate treatment. For present purposes, the key points to recognize are that (i) proxy discrimination can be either intentional or unintentional, and (ii) unintentional proxy discrimination represents one specific type of disparate impact claim.

---

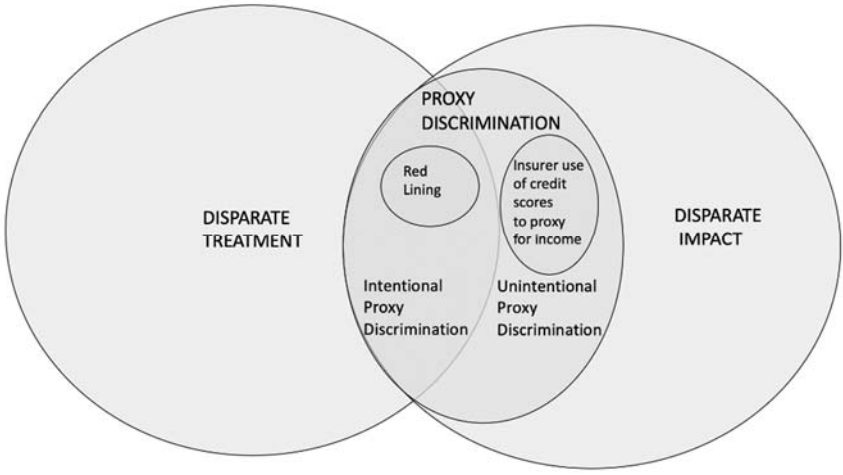
48. *See supra* Section II.A.

49. *See* Steeg Morris, Schwarcz & Teitelbaum, *supra* note 9, at 402–03. Outside of the narrow context of insurance that is linked to housing, disparate impact theories are generally not cognizable in insurance law. And even within the housing setting, the availability of a disparate impact cause of action under the Fair Housing Act is unclear, turning on complex issues of "reverse-preemption" under the McCarran Ferguson Act. *See* KENNETH S. ABRAHAM & DANIEL SCHWARCZ, INSURANCE LAW AND REGULATION 151–57 (6th ed. 2015).

50. Indeed, some commentators have suggested that a primary purpose of disparate impact is to target intentional discrimination that is too difficult to prove. *See, e.g.*, George Rutherglen, *Disparate Impact Under Title VII: An Objective Theory of Discrimination*, 73 VA. L. REV. 1297, 1297–98 (1987) (describing the difficulties and "ambiguities surrounding the theory of disparate impact," which have "obscured the differences between disparate impact and disparate treatment" and led to confusion).

51. Griggs v. Duke Power Co., 401 U.S. 424, 429–31 (1971) (holding that "practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to 'freeze' the status quo of prior discriminatory employment practices"), *superseded by statute*, Civil Rights Act ("Title VII") of 1964, 42 U.S.C. § 2000e-2, *as recognized in* U.S. v. State of North Carolina, 914 F. Supp. 1257 (E.D.N.C. 1996).

Figure 1



B. PROXY DISCRIMINATION BY AIS

Big data and AI are game changers when it comes to the risk of unintentional proxy discrimination. In particular, proxy discrimination by AIs is virtually inevitable whenever the law seeks to prohibit use of characteristics whose predictive power cannot be measured more directly by facially neutral data ("directly predictive characteristics").

Appreciating this point requires a rudimentary understanding of how AIs generate predictions using big data. Such machine learning "automates the process of discovering useful patterns" between characteristics and desired outcomes.[52] To do so, a computer program (the AI) is first "trained" on a dataset for which the outcome of interest, known as the target variable, is known.[53] For instance, the AI might be trained on data for preexisting policyholders, which includes both (i) data on past and existing customers (input data), and (ii) the outcome of interest for these policyholders, such as ultimate claims payouts (target variable).

The scale of such training data has increased dramatically in recent years. Traditionally, firms differentiated among customers, employees, and others based on a limited amount of data that they directly collected. In recent years, however, firms have increasingly come to rely on data secured from a broad number of external sources. These data frequently involve online actions, such as "transactions, email, video, images, clickstream, logs, search queries,

52. Barocas & Selbst, *Big Data*, *supra* note 5, at 677 (examining the concerns that arise from using data mining to remove human biases from the decision making process). *See generally* Gillis & Spiess, *supra* note 29 (analyzing current legal requirements with the structure of AI to identify the issues between old law and new methods).

53. Barocas & Selbst, *Big Data*, *supra* note 5, at 677–78.

health records, and social networking interactions . . . ."[54] But firms rely on data that increasingly also extends to actions in the physical world, which are measured by "sensors deployed in infrastructure such as communications networks, electric grids, global positioning satellites, roads and bridges, as well as in homes, clothing, and mobile phones."[55]

From this training data, the AI derives complex statistical models linking the input data with which it has been provided to predictions about the target variable.[56] In doing so, the AI entirely ignores potential explanations for these relationships, which are immaterial to its programmed goal of maximizing or minimizing the desired outcome, such as aggregate predicted claims expenses.[57] And unlike traditional statistical models, the AI does not start from any overarching theory or hypothesis regarding what types of characteristics may prove useful for predicting the target variable.[58] Instead, the AI effectively uses brute force to "learn" which attributes or activities predict the outcome of interest.[59] For this reason, the ultimate statistical models that AIs derive are often nearly impossible to explain intuitively; the models work, but no one—including the programmer, the firm that relies on it, or the AI itself—can explain *why* or *how* it does so.[60]

As a computer program, of course, AIs do not have any conscious awareness or objectives that are independent from those that are embedded within their code. For this reason, most commentators and courts believe that an AI cannot itself engage in intentional discrimination, at least apart from its programmer or user.[61] Although some have suggested that algorithmic decision-making could, and should, be conceptualized as intentional discrimination, adjudication of this debate is beyond the scope of this paper.[62]

---

54. Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 240 (2013).

55. *Id.*

56. *See* O'NEIL, *supra* note 5.

57. Crawford & Schultz, *supra* note 5, at 99; *see also* Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 405 (2017); Rick Swedloff, *The New Regulatory Imperative for Insurance*, 61 B.C. L. REV. (forthcoming 2020) (manuscript at 6).

58. COMM. ON TECH., EXEC. OFFICE OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE 8 (2016), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf; Matthew T. Bodie, Miriam A. Cherry, Marcia L. McCormick & Jintong Tang, *The Law and Policy of People Analytics*, 88 U. COLO. L. REV. 961, 969–71 (2017); Allan G. King & Marko J. Mrkonich, *"Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555, 555 (2016).

59. Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 15 (2019) ("The algorithm itself tries many possible combinations of variables, figuring out how to put them together to optimize the objective function.").

60. *See* Bruckner, *supra* note 4, at 44–46; Coglianese & Lehr, *supra* note 59. There is a substantial computer science movement that is working on developing AIs that can explain their outputs. *See, e.g.*, Turek, *supra* note 14; *Making Computers Explain Themselves*, *supra* note 14.

61. *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 699.

62. *See* Bornstein, *supra* note 7, at 571.

Consistent with the prevailing view, we assume that all forms discrimination by AI cannot be intentional unless some person intentionally embeds within the AI an illicit discriminatory objective or methodology, or at the very least is aware that the AI is acting in a discriminatory fashion and continues to employ the algorithm.

Armed with this basic understanding of AI and big data, it is now possible to understand why these forces will inevitably produce unintentional proxy discrimination when the law seeks to prohibit discrimination based on directly predictive characteristics.[63] This conclusion follows inevitably from the nature of predictive AIs, which are directly programmed to find linkages between input data and target variables, irrespective of the nature of these linkages. By using the data it is trained on to proxy for directly predictive but legally suspect information, AIs optimize their programmed objective. Moreover, as they are provided with more and more training data, they will become better and better at identifying proxies for directly predictive, but legally prohibited, characteristics.[64]

This unintentional proxy discrimination by AIs cannot be avoided merely by depriving the AI of information on individuals' membership in legally suspect classes or obvious proxies for such group membership.[65] To be sure, this traditional approach to anti-discrimination law may prevent intentional proxy discrimination by human actors. However, it fails in the context of unintentional proxy discrimination by AIs, because AIs can and will use training data to derive less intuitive proxies for directly predictive characteristics when they are deprived of direct data on these characteristics due to legal prohibitions.[66]

These conclusions are consistent with the emerging consensus in the extant literature that simply depriving AIs of direct data on protected characteristics does not necessarily prevent those algorithms from exhibiting bias.[67] But the point here is more specific to proxy discrimination; depriving

63. *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 712 (illustrating that at least one other commentator has briefly suggested parallels between redlining and statistical proxy discrimination by AIs); *see also* Sullivan, *supra* note 15, at 416 ("In still pursuing good employees, [perhaps] the most likely scenario is that Arti will use proxies for the forbidden traits (second-best criteria) to achieve results that approximate what it would have done had not sex been ruled out-of-bounds. If a human were to undertake this exercise, we might well talk of 'masking' her true motive, but we've seen that Arti has no motive[s]." (footnotes omitted)).

64. *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 695 (An AI armed "with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait whether or not that information is an input.").

65. For further discussion of this point, see *infra* Section IV.A.

66. *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 691–92.

67. *See, e.g.*, Gillis & Spiess, *supra* note 29, at 464 ("However, the exclusion of the forbidden input alone may be insufficient when there are other characteristics that are correlated with the forbidden input—an issue that is exacerbated in the context of big data."); Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Ashesh Rambachan, *Algorithmic Fairness*, 108 AEA PAPERS & PROC. 22, 22 (2018) ("Numerous studies (many of them in computer science) have pointed out that this

algorithms of directly predictive but suspect characteristics does not merely leave open the possibility of algorithms exhibiting various biases. Instead, this strategy will *inevitably* fail to prevent proxy discrimination based on suspect characteristics that are directly predictive of the target variable. This is because increasingly sophisticated AIs will affirmatively "seek out" proxies for prohibited, but predictive, characteristics within increasingly vast amounts of training data. To illustrate, an AI deprived of information about a person's genetic test results or obvious proxies for this information (like family history) will use other information—ranging from TV viewing habits to spending habits to geolocational data—to proxy for the directly predictive information contained within the genetic test results.

AI and big data, in sum, are poised to take the problem of unintentional proxy discrimination from a niche and under-theorized issue into a pervasive concern for all antidiscrimination regimes that seek to limit the use of protected traits that are directly predictive. But understanding the scale and urgency of this shift requires disentangling several different scenarios when legally suspect characteristics may be directly predictive of legitimate outcomes. We now turn to this task.

C. UNDERSTANDING WHEN PROXY DISCRIMINATION BY AIS IS LIKELY TO OCCUR

1. Direct and Indirect Proxy Discrimination

In an ideal setting, employers, insurers, lenders, and other social actors would isolate the underlying causes of their desired outcomes and differentiate solely on these bases. Do aggressive driving patterns—as recorded by telematic equipment or other GPS enabled devices—cause more auto insurance claims? If so, then insurers could simply reduce expected insurance claims by choosing to insure those with less aggressive driving patterns. Of course, the causes of future states of the world are rarely fully known or understood, a reality that AI and machine learning do little to alter. Instead, these technologies focus solely on identifying correlations between known variables, on the one hand, and desired future states of the world, on the other.[68]

---

requires more than just excluding race from the predictor, since protected features such as race could be reconstructed from other features.").

68. Some may argue that actors do not actually care about causation. For example, the argument goes, why would an insurer care whether a bad credit score causes a life insurance claim or not—as long as they can lower the riskiness of their insurance pool, who cares what is causative and what is correlative? Practically, this is true since determining causation is rarely a possibility. However, relying on correlation will naturally leave error in the risk pool. There will be some with poor credit scores who will live a long life. In the arms race of underwriting, the first insurer to determine how to best split those with poor credit scores into those with low credit who are at risk of early death versus those with low credit who are not will have the upper hand against other insurers. If these insurers knew true causation, they would have an accurate assessment of the riskiness in their pool. This could also break insurance, but that is neither here nor there for this Article.

However, the risk that AIs will in fact proxy discriminate depends substantially on the different pathways of causation and correlation that link a legally protected characteristic and a target variable of interest. To the extent that there is no such link—as is undoubtedly the case in many scenarios—then proxy discrimination by AI is not possible, even if disparate impact may be. By contrast, as described above, AIs will inevitably tend to proxy discriminate whenever the law prohibits discrimination on the basis of a directly predictive characteristic, meaning that the characteristic's power to predict a desired "target variable" cannot be captured more directly by facially neutral data. There are two different ways in which this condition can be met, which we label casual and opaque proxy discrimination. Proxy discrimination is also possible when a protected characteristic has predictive power solely because it correlates with a known, facially-neutral characteristic. As explained below, we label this indirect proxy discrimination.

Causal Proxy Discrimination (Direct)

Variable (X)	Proxies for (Y)	Which causally predicts (Z)
Facially neutral classifier	Suspect classifier	Desired outcome

First, a legally-suspect characteristic can be directly predictive because it is causally linked to the desired outcome, as depicted above.[69] For present purposes, a suspect classifier is causally predictive of some future state of the world when its presence would always impact the probability of the targeted outcome in a statistical model, irrespective of any additional information that could be added to that model.[70] In other words, causation requires a direct link between a suspect classifier and a desired outcome such that the predictive power of the suspect classifier is not itself a result of it proxying for some omitted or unknown characteristic. The desired outcome can encompass a variety of measures, from end-goal characteristics, such as likelihood of filing claims or defaulting on a loan, to market-based outcomes, such as price elasticity or likelihood to stay in one place of employment for multiple years.

Perhaps the best example of such a causally predictive characteristic is the gene for Huntington's disease, which is essentially 100 percent penetrant —individuals with a series of nucleotide repeats in the *HTT* gene over a set threshold are essentially always going to develop the disease, whereas those

69. Max N. Helveston, *Consumer Protection in the Age of Big Data*, 93 WASH. U. L. REV. 859, 866 (2016).

70. *See generally* JUDEA PEARL & DANA MACKENZIE, THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT (2018) (developing a general theory of causation that focuses on counterfactual questions regarding what would occur in various hypothetical scenarios).

below a certain threshold will never develop the disease.[71] There are no other variables, such as environmental causes or other genes, that help to predict whether Huntington's disease will develop.[72] Direct causal relationships like these are hard to isolate. But when they exist, sufficiently sophisticated AIs deprived of direct information about these characteristics due to legal restrictions will identify and use any available data that even partially proxies for this information. For instance, in the case of Huntington's disease, if algorithms were legally prohibited from taking into account genetic tests for the disease (Y), they could proxy for the disease through variables (X) like family medical history or visits to a website for a Huntington's disease support group.[73]

Opaque Proxy Discrimination (Direct)

Variable (X)	Proxies for (Y)	Proxies for (A)	Which causally predicts (Z)
Facially neutral classifier	Suspect classifier	Unquantifiable or unavailable variable	Desired outcome

A second scenario in which a legally suspect characteristic can be directly predictive—thus tending to produce proxy discrimination by AIs—is when it is correlated to a desired outcome, but its predictive character is not mediated through a presently quantifiable or available variable. We label this opaque proxy discrimination.

Opaque proxy discrimination can occur in two scenarios. First, it may be that the causative variable for which the suspect classifier is proxying cannot be quantified because it is not fully understood. If so, then it may be that the suspect variable is in fact causative or that it is merely proxying for a causative factor. For example, in the genetics context, even for many pathogenic genetic variants, it is often unknown why a particular sequence in a gene leads to increased risk.[74] It may well be that one gene has been identified as higher risk because it is correlated with some other more particular DNA segment that has yet to be identified and characterized. Alternatively, it may be that the true causative mechanism is epigenetic changes that turn on and off the gene in question.

71. *Huntington Disease*, U.S. NAT'L LIBR. MED., https://ghr.nlm.nih.gov/condition/huntington-disease#genes [https://perma.cc/NWB3-QEPH] (last reviewed June 2013).

72. *Id.*

73. For Huntington's disease, there is a 50 percent chance of inheriting the genetic marker, and thus developing the disease, if a parent had Huntington's. *Id.* If a grandparent had Huntington's, but it is not known whether the parent did, the chance of developing the disease is 25 percent. *Id.*

74. *See* Brendan Bulik-Sullivan et al., *An Atlas of Genetic Correlations Across Human Diseases and Traits*, 47 NATURE GENETICS 1236, 1236 (2015).

The other scenario in which opaque discrimination can occur is when the suspect variable proxies for a true causative variable that is understood, but nonetheless difficult to quantify. A good example involves sex and auto insurance. Sex (Y) is predictive of auto insurance claims (Z) in part because young girls tend to drive more safely than young boys.[75] Of course, it is possible to obtain more direct information about care levels (A). But such data is not widely available, as driver "care" is difficult to quantify. For this reason, simply banning the use of sex-based discrimination will predictably lead to proxy discrimination by AIs because sex is directly predictive of care levels in ways that are not mediated through any alternative, presently quantifiable, variables.

Proxy discrimination by AIs is just as likely to occur when the link between the suspect variable and target variable is opaque as compared to when it is causal. In both cases, the suspect variable is "directly predictive." But unlike in the case of causal proxy discrimination, AIs engaging in opaque proxy discrimination may cease to proxy discriminate in the future if new facially neutral data becomes available that more directly proxies for the true causative variable than the suspect variable. Returning to the example of sex and auto insurance, insurers are increasingly generating more direct data about driver care levels through techniques like telematics. As this data becomes more widely available, AIs may shift from proxy discriminating based on sex to discriminating based on non-suspect and more direct measures of driver care, like frequency of sudden stops.

Indirect Proxy Discrimination

Variable (X)	Proxies for (Y)	Proxies for (A)	Which causally predicts (Z)
Facially neutral classifier	Suspect classifier	Quantifiable and available variable	Desired outcome

In both causal and opaque proxy discrimination, prohibited characteristics are "directly predictive" of legitimate outcomes of interest. But proxy discrimination may also occur due to indirect connections between prohibited traits and target variables. In particular, proxy discrimination will tend to occur when a suspect variable is predictive of a desired outcome only because it proxies for another, quantifiable and potentially available, variable

75. *See Rating Automobile Insurance: Testimony Before the Subcomm. on Oversight & Investigations of the H. Comm. on Fin. Servs.*, 116th Cong. 4–5 (2019) (statement of James Lynch, Chief Actuary and Senior Vice President of Research and Education, Insurance Information Institute), *available at* https://financialservices.house.gov/uploadedfiles/hhrg-116-ba09-wstate-lynchj-20190501.pdf [https://perma.cc/YET5-AFRZ].

that causes the desired outcome but that is not included in the AI's training data.[76] This type of indirect proxy discrimination is visually depicted above. In these cases, the predictive power of the original facially neutral classifier is attributable to its correlation with the suspect classifier, whose predictive power is, in turn, attributable to its correlation with the causative facially neutral characteristic. The suspect variable does not itself constitute "directly predictive" data in these cases; instead, it is predictive merely because it provides one of several potential ways to assess the likelihood of some true causative factor that is both quantifiable and potentially available, but not in fact accessible to the AI.

This is akin to omitted variable bias in statistics. For instance, height (A) might be directly predictive of job performance for job (Z), but the AI might lack access to data on the current applicants' heights. In this case, the algorithm may find that applicants' sex (Y) is an imperfect proxy for height (A), since height and sex are highly correlated. Deprived of information on sex due to laws prohibiting discrimination on this basis, the algorithm might use a proxy for sex, such as applicants' Netflix viewing habits (X), to predict the outcome of interest (Z). In that sense, indirect proxy discrimination is the AI parallel to "statistical discrimination"; the AI would be acting just as an employer who refuses to interview people with traditionally female first names because there is a legitimate job-specific reason for hiring tall employees and height is not specified on job applicants' resumes.[77]

Unlike both causal and opaque proxy discrimination—where the suspect variable is directly predictive—indirect proxy discrimination is simply a possible, but hardly inevitable, result of algorithms. Indirect proxy discrimination will not occur if either data on the causative facially neutral characteristic (A) is included in the model directly, or if better proxies than the suspect characteristic are available to the AI.[78] Returning to the example of sex discrimination and height, an AI will not engage in indirect proxy discrimination if it can directly access data on height (a non-suspect variable) or can proxy for height more effectively by exclusively relying on factors that are not linked to sex, like recent clothing purchases.

For these reasons, indirect proxy discrimination may well tend to decrease as more data is added into training data and AIs become more sophisticated. However, if new data becomes available but is not incorporated into a particular AI, preexisting proxy discrimination will continue.

---

76. Barocas and Selbst describe this as "rational racism." Barocas & Selbst, *Big Data*, *supra* note 5, at 690. "Accordingly, the persistence of distasteful forms of discrimination may be the result of a lack of information, rather than a continued taste for discrimination." *Id.*

77. *See* Agan & Starr, *supra* note 42, at 193–94.

78. *See* W. Nicholson Price II, Note, *Patenting Race: The Problems of Ethnic Genetic Testing Patents*, 8 COLUM. SCI. & TECH. L. REV. 119, 134–37 (2007) (discussing why race is a poor proxy in genetic tests since those tests directly evaluate the underlying trait relevant to the outcome).

Analytically, such proxy discrimination would shift from opaque proxy discrimination to indirect proxy discrimination.

A summary of these three types of proxy discrimination—causal, opaque, and indirect—is contained in Figure 2, below.

Figure 2

Types of proxy discrimination	Definition	Is Suspect variable Directly or Indirectly predictive?	Risk of Proxy Discrimination by AI	Examples
Causal proxy discrimination	Legally-suspect characteristic (i.e. race, genetics, health) causally linked to target variable (i.e. expected insurance costs).	Directly Predictive, as suspect variable contains predictive power that cannot be more directly captured by facially-neutral data.	Very high risk as AIs will inevitably proxy for suspect characteristic.	GINA prohibition on genetic information in employment and health insurance.
Opaque proxy discrimination	Legally-suspect characteristic (i.e. race, genetics, health) predictive of target variable (i.e. expected insurance costs) for reasons not mediated through a presently quantifiable or available variable.	Directly Predictive, as suspect variable contains predictive power that cannot be more directly captured by facially-neutral data.	High risk as AIs will inevitably proxy for suspect characteristic until better data or causal mechanism becomes available.	Auto state insurance prohibition on use of sex to proxy for driver care (which is not readily quantifiable).
Indirect proxy discrimination	Legally-suspect characteristic (i.e. race, genetics, health) predictive of target variable (i.e. expected insurance costs) because it proxies for a quantifiable or available variable.	Indirectly Predictive, as suspect variable only contains predictive power because it proxies for another, quantifiable and potentially available, variable that is not included in the AI's training data.	Moderate risk as AIs will only proxy discriminate if (i) data on causative facially-neutral characteristic is not available, and (ii) better proxies for causative characteristic than suspect characteristic are not available.	Auto state insurance prohibition on use of sex to proxy for miles driven (which is quantifiable and potentially available).

2. The Difficulty of Identifying Causal, Opaque, and Indirect Proxy Discrimination by AIs in the Real World

While it is helpful to parse out each of these potential types of proxy discrimination—causal, opaque, and indirect—in reality, the predictive value of the myriad available variables in a big data world is much more complex. For every algorithmic prediction of a desired outcome there is: usually more than one explanation; evidence of correlation, not causation; and voluminous amounts of data to explore. As such, identifying ahead of time how likely a particular AI is to proxy discriminate is an immensely difficult, if not impossible, task.

Rarely is there just one causative explanation for a desired outcome. Rather, multiple variables or combinations of variables predict an outcome.[79]

79. Pope & Sydnor, *supra* note 9, at 206. *See generally* Bruce Glymour & Jonathan Herington, *Measuring the Biases that Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms*, PROCEEDINGS OF ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY AND TRANSPARENCY 269, 270 (2019) (mapping a variety of potential causal mechanisms possible in a model).

For this reason, different types of proxy discrimination are likely to all occur at the same time. Take sex and mortality. In general, women live longer than men—there is a correlation between sex and life expectancy.[80] Thus, where life insurers are prohibited from directly considering sex, the algorithms they employ may rely on facially-neutral proxies for sex. But the predictive power of sex almost certainly derives substantially from its capacity to proxy for other omitted and measurable variables, such as utilization of healthcare, workplace exposures and hardships, and risky behaviors like drinking, smoking, and overeating.[81] Additionally, there are probably some socio-cultural contributors to life expectancy that cannot be readily measured, such as how one's role as caregiver could increase self-esteem and recognition in a way that leads to longer lives.[82] Finally, there are some biological differences between the sexes that causally explain variance in life expectancy, such as differences in hormones.[83]

Thus, all three types of proxy discrimination are at play when an AI proxies for sex in predicting life expectancy by, for instance, using social media likes or names to proxy for sex. First, the algorithm is engaging in causal unintentional proxy discrimination, as biological sex has a causal explanation for some elements of life expectancy. Second, sex is standing in as a proxy for other unknown or unmeasurable variables, implicating opaque proxy discrimination. Third, because sex is itself a proxy variable for omitted facially neutral variables, such as how much one smokes, the AI is engaging in indirect proxy discrimination. All three types of proxy discrimination exist in the same correlative relationship because, in reality, there is rarely one distinct cause of a desired outcome—most variables are not like the gene for Huntington's disease. Each partially predicts the desired outcome.

Not only are all three types of proxies likely to appear in the same model, they will often build on each other. For example, a suspect classifier (age) may proxy for a facially neutral category (years since graduation) which proxies for some unquantifiable data (comfort with learning new technology), which predicts a desired outcome. Alternatively, an AI may proxy for one suspect classifier, which proxies for another suspect classifier, which proxies for a facially neutral characteristic that is casually linked to the target variable. To illustrate this possibility, reconsider the height and sex example above, where an AI proxies for sex though a facially neutral variable (such as shopping

80. Bertrand Desjardins, *Why Is Life Expectancy Longer for Women than It Is for Men?*, SCI. AM. (Aug. 30, 2004), https://www.scientificamerican.com/article/why-is-life-expectancy-lo [https://perma.cc/ZE3B-B8AX].

81. *Id.*

82. *Johns Hopkins-Led Study Shows Increased Life Expectancy Among Family Caregivers: Findings Contradict Long-standing Beliefs About Caregiver Stress*, JOHNS HOPKINS MED. (Oct. 15, 2013), https://www.hopkinsmedicine.org/news/media/releases/johns_hopkins_led_study_shows_increased_life_expectancy_among_family_caregivers [https://perma.cc/NV4U-QGUW].

83. Desjardins, *supra* note 80.

patterns) because sex proxies for height, which is relevant to job performance. In this example, the data on shopping patterns may in fact proxy for gender (a suspect characteristic), which in turn proxies for sex (also a suspect characteristic), which ultimately proxies for the facially neutral characteristic of height.

The upshot of these complexities is that while it is relatively easy in theory to identify when an AI is likely to engage in proxy discrimination, it is immensely difficult to do so in practice. Of course, this is only a problem if proxy discrimination by AIs is itself troubling from a broader social perspective. As we explore in the next Part, this is undoubtedly the case.

III. THE HARMS OF PROXY DISCRIMINATION BY AIS

When an AI proxy discriminates, it uses a facially neutral variable to capture the predictive power of a legally prohibited trait. This Part explores the potential implications of such proxy discrimination by AI. To do so, it first identifies the many different settings in which anti-discrimination laws do, in fact, prohibit discrimination on the basis of traits that are directly predictive of discriminators' legitimate goals. In these circumstances, AIs will tend to capture the prohibited trait's predictive power using facially neutral data proxies, as discussed above, unless the law affirmatively prevents this outcome from obtaining, a possibility we discuss in Part IV.

Second, this Part explains why proxy discrimination by AIs is so troubling from a normative perspective. Ultimately, the argument is straight-forward: Laws that prohibit discrimination based on directly predictive traits are normatively grounded in the goal of preventing specific outcomes for members of protected groups. Unlike some anti-discrimination settings, the questions of how or why bad outcomes are experienced by protected groups are secondary, if relevant at all, in these domains. Because proxy discrimination by AIs tends to produce the very same outcomes that would result in the absence of legal restrictions on discrimination based on directly predictive traits, it represents a substantial threat to the normative underpinnings of these anti-discrimination regimes.

A. *ANTI-DISCRIMINATION REGIMES AT RISK OF PROXY DISCRIMINATION BY AIS*

Not infrequently, discrimination based on a legally suspect trait is rational because the trait contains predictive power that cannot be more directly captured by available facially neutral data. In other words, the data is "directly predictive" of the outcome of interest.[84] The law nonetheless bars actors from taking into account these traits because doing so has broader normative implications.[85] As suggested in Part II, proxy discrimination by AIs

84. *See supra* Section II.C.1.

85. *See generally* Samuel R. Bagenstos, *"Rational Discrimination," Accommodation, and the Politics of (Disability) Civil Rights*, 89 VA. L. REV. 825 (2003) (taking a normative approach to the assertion

is a substantial, and nearly inevitable, risk in these settings, at least absent affirmative counteracting legal strategies like those we discuss in Part IV. These initial conditions where proxy discrimination by AIs is likely to flourish are most obvious in insurance, but also exist in other settings, such as employment and education.

1. Health Insurance

Numerous state and federal laws prohibit health insurers from discriminating on the basis of directly predictive characteristics. Most notably, the ACA[86] prohibits or limits discrimination on the basis of prior health history, preexisting conditions, age, sex, and smoking history.[87] Indeed health insurers are currently only able to consider up to four traits when setting insurance premiums.[88] Many individual states also prohibit discrimination based on some, or all, of these individual traits.[89] Each of these legally-suspect characteristics are, of course, directly predictive of health insurers' expected claims expenses, as they predict future medical expenses for reasons that cannot be more directly captured by alternative, facially-neutral data.

The ACA is by no means the only law that bars health insurers from discriminating on the basis of directly predictive, and potentially causal, information. In particular, GINA bars health insurers and employers from discriminating on the basis of genetic test results or several obvious proxies

---

that the effects of accommodation requirements are similar to those of antidiscrimination requirements). Of course, in many cases the membership in a protected class will be "irrelevant to the outcome in terms of discriminatory effect, at least given a large number of input features." *See* Barocas & Selbst, *Big Data*, *supra* note 5, at 695.

86. *See* Individual and Group Market Reforms, 42 U.S.C. §§ 300gg–300gg-2 (2012). Some "health insurance" plans like short duration plans or association plans are not required to comply with the ACA and therefore retain the ability to underwrite on a broad set of traits. *Id.*

87. Under the ACA, insurers can vary rates based on only four factors: (1) whether a plan covers an individual or family; (2) a "rating area" or geographic area designated by the state; (3) age; and (4) smoking status. *Id.* § 300gg(a)(1)(A). Even the use of these characteristics is constrained, as the law sets allowable ratios across subgroups of individuals with the characteristic. *Id.* § 300gg(a)(1)(A)(iii)–(iv). These restrictions on ratemaking are coupled with guaranteed issue and renewability provisions that require insurers to accept all applications for health insurance and to continue to insure existing policyholders as long as they pay premiums. *Id.* §§ 300gg-1–300gg-2. Additionally, the ACA explicitly prohibits several types of rational discrimination, most notably the use of gender and pre-existing conditions. *See e.g.*, Sherry A. Glied & Adlan Jackson, *Access to Coverage and Care for People with Preexisting Conditions: How Has It Changed Under the ACA?*, COMMONWEALTH FUND (June 22, 2017), https://www.commonwealthfund.org/publications/issue-briefs/2017/jun/access-coverage-and-care-people-preexisting-conditions-how-has [https://perma.cc/MXK4-58KX].

88. 42 U.S.C. § 300gg(a)(1)(A); *see infra* Part IV.

89. Some states have further restricted allowable ratios for age and smoking—sometimes all the way down to 1:1, thus essentially removing the characteristic from consideration. CTR. FOR CONSUMER INFO. & INS. OVERSIGHT, CTRS. FOR MEDICARE & MEDICAID SERVS., *Market Rating Reforms: State Specific Rating Variations*, https://www.cms.gov/CCIIO/Programs-and-Initiatives/Health-Insurance-Market-Reforms/state-rating.html [https://perma.cc/8TM8-FDBS] (last updated June 2, 2017).

for such information.[90] Yet there is little doubt that certain types of genetic test results are—or will be in the future—directly predictive of legitimate considerations for both employers and insurers.[91] For example, genetic tests for early onset Alzheimer's could help employers or health insurers identify individuals at increased risk of needing costly healthcare interventions. As genetic information becomes better understood and more widely accessible, this possibility that genetic information may be directly predictive will only increase. As Part IV discusses, the ACA, but not GINA, partially limits the possibility of proxy discrimination.

2. Non-Health Insurance

Unlike health insurers, non-health insurers such as life, automobile, property, or disability insurers are regulated predominantly by the states. And under state laws prohibiting "unfair discrimination," these insurers can generally discriminate on the basis of traits if, and only if, they are predictive of risk.[92] But there are also important legal prohibitions on specific types of discrimination by non-health insurers, though they vary significantly by state and line of insurance.[93] These include prohibitions on insurance discrimination based on: race, national origin, ethnicity, sexual orientation, age, income, credit scores, marital status, disability, length of driving experience, genetic information, and many others.[94] The implication of this structure is that specifically-prohibited traits cannot be used by insurers even if they are predictive of risk; otherwise trait-specific prohibitions in insurance would be superfluous given more general laws banning "unfair discrimination."

Perhaps the most intuitive example of this structure involves state laws prohibiting insurers from discriminating against individuals who have been victims of intimate partner violence. Historically, insurers frequently discriminated against this population precisely because they were genuinely at greater risk of death, injury, or property destruction.[95] Despite the fact that a history of intimate partner violence is directly predictive of insurers' outcome of interest (insurance claims), many states chose to ban such

90. *See* Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233, 122 Stat. 881 (codified as amended in scattered sections of 26, 29, and 42 U.S.C).

91. There is, however, an ongoing debate about the actuarial significance of most information that comes from genetic tests due to the complexities of environment and biological mechanisms of disease. *See generally* Prince, *supra* note 26 (examining the ethical, financial, and legal questions presented by this debate).

92. *See* ABRAHAM & SCHWARCZ, *supra* note 49, at 138; Prince, *supra* note 26, at 640; Schwarcz, *supra* note 16, at 987.

93. *See* Avraham et al., *supra* note 38, at 243; Leah Wortham, *Insurance Classification: Too Important to Be Left to the Actuaries*, 19 U. MICH. J.L. REFORM 349, 387–92 (1986).

94. ABRAHAM & SCHWARCZ, *supra* note 49, at 138–41.

95. *See* Ellen J. Morrison, Note, *Insurance Discrimination Against Battered Women: Proposed Legislative Protections*, 72 IND. L.J. 259, 275 (1996).

discrimination by statute or regulation.[96] A second intuitive example involves state prohibitions of sex-based discrimination in auto insurance.[97] Such discrimination is common in the absence of legal prohibitions precisely because sex is directly predictive of claims, as young women tend to drive more safely than young men and most auto insurers have limited alternative data that more directly predicts safe driving.[98] Nonetheless, several states ban such discrimination.

3. Employment

Employment anti-discrimination law is another important example of a regime that prohibits discrimination based on directly predictive characteristics.[99] As in the health insurance context, GINA is illustrative: Under GINA, employers are prohibited from considering genetic information, even though it could help predict any number of facially legitimate outcomes of interest to employers, such as anticipated productivity or longevity of tenure.

But GINA is hardly an isolated example. The Americans with Disabilities Act ("ADA"),[100] for example, prohibits discrimination against individuals with disabilities who can perform the essential functions of a job with or without "reasonable accommodations."[101] This is true even if the individual's disability may be directly predictive of outcomes like costs spent on accommodations, group health insurance costs, or longevity.[102] The Pregnancy Discrimination

---

96. *See, e.g.*, KY. REV. STAT. ANN. § 304.12-211(2)(a) (LexisNexis 2011); *see also* ABRAHAM & SCHWARCZ, *supra* note 49, at 276. *See generally* Deborah S. Hellman, *Is Actuarially Fair Insurance Fair?: A Case Study in Insuring Battered Women*, 32 HARV. C.R.-C.L. L. REV. 355, 361–69 (1997).

97. *See, e.g.*, Ann Carrns, *In California, Gender Can No Longer Be Considered in Setting Car Insurance Rates*, N.Y. TIMES (Jan. 18, 2019), https://www.nytimes.com/2019/01/18/your-money/car-insurance-gender-california.html [https://perma.cc/96FF-WZFZ].

98. For younger drivers, women tend to have fewer claims than men. Some insurers report that this trend reverses for older drivers, though insurers have different experiences on this point. *See id.*

99. *See* Sullivan, *supra* note 15, at 402–03; *see also* Barocas & Selbst, *Big Data*, *supra* note 5, at 694–713 (detailing employment antidiscrimination frameworks).

100. Americans with Disabilities Act of 1990, Pub. L. No. 101-336, § 2(b), 104 Stat. 327, 329 (codified as amended at 42 U.S.C. §§ 12101–12213 (2012)).

101. *See* 42 U.S.C. § 12102; *see also* Samuel Issacharoff & Justin Nelson, *Discrimination with a Difference: Can Employment Discrimination Law Accommodate the Americans with Disabilities Act?*, 79 N.C. L. REV. 307, 314–15 (2001) (arguing that the ADA treats "differently situated" persons differently in its reasonable accommodation standard, unlike other employment laws).

102. *See generally*, Mark Kelman, *Market Discrimination and Groups*, 53 STAN. L. REV. 833 (2001) (discussing reasonable accommodations required by the ADA). *See* Bagenstos, *supra* note 85, at 832 ("[A]ccommodation requirements represent nothing more than a specific example of the general prohibition of rational discrimination—a prohibition that is well entrenched in the law."); *see also* Sharona Hoffman, *Big Data's New Discrimination Threats: Amending the Americans with Disabilities Act to Cover Discrimination Based on Data-Driven Predictions of Future Disease*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS 85, 85–87 (I. Glenn Cohen et al. eds., 2018) (examining how

Act ("PDA")[103] also prohibits discrimination "on the basis of pregnancy, childbirth, or related medical conditions."[104] Furthermore, it requires that employers must provide reasonable accommodations to women who are temporarily disabled due to pregnancy.[105] Pregnancy—or factors suggesting the likelihood of future pregnancy—would likely be directly predictive of facially neutral objectives for many employers, most obviously the likelihood of a prospective employee taking an extended leave of absence.[106]

Yet another intuitive example of an employment law that proscribes directly predictive discrimination is the Age Discrimination in Employment Act of 1967 ("ADEA"),[107] which prohibits discrimination against individuals who are 40 years of age or older.[108] Age is almost certainly predictive of at least some employers' expected returns on prospective employees.[109] Older job applicants generally have fewer remaining working years than younger applicants, which may limit the extent to which they are likely to advance within the organization.[110] Older employees may also be more likely to take medical leaves than younger workers due to health complications.[111] For these reasons, AIs may well proxy discriminate for age when producing hiring or advancement recommendations for a variety of employers.

Finally, Title VII bars employers from discriminating on the basis of race, color, religion, sex and national origin.[112] Although these traits are less intuitively 'directly predictive' of outcomes of interest, they can, in fact, meet this condition. For example, race is correlated with a wide variety of outcomes

---

employers may also be interested in using big data to identify those who are predicted to get a disability in the future; however, noting that such predictive health information is not adequately legally protected in anti-discrimination laws). Indeed, the reasonable accommodation features of the ADA led to a wide-ranging legal literature regarding the extent to which the law paralleled more conventional federal anti-discrimination regimes, like Title VII of the Civil Rights Act of 1964. *See, e.g.*, Christine Jolls, Commentary, *Antidiscrimination and Accommodation*, 115 HARV. L. REV. 642, 644, 651–52, 672–74 (2001) (describing the distinction between accommodation requirements and antidiscrimination laws); Pamela S. Karlan & George Rutherglen, *Disabilities, Discrimination, and Reasonable Accommodation*, 46 DUKE L.J. 1, 2–3 (1996).

103. Civil Rights Act ("Title VII") of 1964, 42 U.S.C. § 2000e(k) (2012).

104. *Id. See also generally* Reva B. Siegel, Note, *Employment Equality under the Pregnancy Discrimination Act of 1978*, 94 YALE L.J. 929 (1985) (providing an overview of the Pregnancy Discrimination Act's mechanism for equalizing the social status of each sex).

105. 42 U.S.C. § 2000e(k).

106. *See* Richard A. Posner, *An Economic Analysis of Sex Discrimination Laws*, 56 U. CHI. L. REV. 1311, 1332–34 (1989) ("[T]he [PDA] compels the employer to ignore a real difference in the average cost of male and female employees.").

107. Age Discrimination in Employment Act of 1996, 29 U.S.C. §§ 621–634 (2012).

108. *See id.* § 631.

109. *See* Steven J. Kaminshine, *The Cost of Older Workers, Disparate Impact, and the Age Discrimination in Employment Act*, 42 FLA. L. REV. 229, 231 (1990) ("[O]lder workers may in fact create costs for employers in ways not encountered under Title VII.").

110. *See id.* at 251.

111. *See id.* at 289.

112. Sullivan, *supra* note 15, at 403.

in American society, from education, to incarceration, to income. Although these correlations generally can be explained by a variety of facially neutral factors, these factors are not always susceptible to direct, quantitative measurement. For that reason, race, unfortunately, is likely to remain directly predictive of a wide range of facially legitimate considerations for many discriminators.[113]

In contexts where explicit or implicit discrimination preexists, traits like race, ethnicity and sex may even be directly predictive because they are causally linked to facially neutral objectives. Consider an example: Amazon recently was forced to abandon an AI that it had developed to identify promising employees, because the AI tended to select male applicants using proxies for sex on applicants' resumes.[114] One likely explanation for this tendency of the AI was that male employees at Amazon had, in fact, been more productive than their female counterparts due to the company's culture implicitly or explicitly favoring men. If so, then sex would be causally linked to the outcome of interest, notwithstanding that consideration of that trait is legally proscribed.

4. Other Legal Areas

Even outside the insurance and employment contexts, the law regularly seeks to prohibit actors from taking into account traits that are directly predictive of an outcome of interest. This is all the more likely as algorithms and big data are increasingly used to make decisions in domains like housing,[115] lending,[116] and policing.

To illustrate, race remains highly predictive of criminal recidivism rates for a variety of difficult-to-quantify reasons.[117] As such, AIs that are programmed to calculate recidivism rates will inevitably seek to capture the predictive power of race by relying on proxies for that characteristic.[118]

---

113. *See* BARADARAN, *supra* note 37 (describing "black banking" and similar initiatives as a decoy for avoiding broader social reforms).

114. *See* Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, REUTERS (Oct. 9, 2018, 10:12 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [perma.cc/KLL3-J3TK].

115. James A. Allen, *The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining*, 46 FORDHAM URB. L.J. 219, 234–35 (2019).

116. *See* Inge Graef, *Algorithms and Fairness: What Role for Competition Law in Targeting Price Discrimination Towards End Consumers*, 24 COLUM. J. EUR. L. 541, 542 (2018) (highlighting how European anti-discrimination laws do not adequately address algorithmic price discrimination concerns); King & Mrkonich, *supra* note 58, at 559; Odinet, *supra* note 4, at 804.

117. *See* Huq, *supra* note 4, at 1047–48; Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 686 (2016).

118. *See, e.g.*, Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 153 (2017) ("[W]e show that the . . . evidence of racial bias in [recidivism prediction models] are a direct consequence of applying an [RPI that satisfies predictive parity] to a population in which recidivism prevalence differs across groups.").

Moreover, in many cases race is predictive of considerations that might matter to a discriminator for directly causative reasons that cannot be disentangled from race itself. For instance, a well-developed literature suggests that minority students often perform less well in educational settings than similarly situated non-minority students, in part, because of "stereotype threat," a phenomenon whereby members of stereotyped groups experience self-fulfilling anxiety/pressure about confirming perceived stereotypes.[119] As such, an AI that a college used to predict prospective applicants' academic performance would likely proxy discriminate for race in order to capture its predictive power. These examples could be replicated for gender, age, disability, and a host of other traits that social actors are commonly barred from taking into account.

B. PROXY DISCRIMINATION BY AIS UNDERMINES THE INTENDED GOALS OF IMPACTED ANTI-DISCRIMINATION REGIMES

Proxy discrimination by AIs is thus a significant risk across a broad spectrum of legal domains that prohibit discrimination based on directly predictive characteristics. But why, one might wonder, is this a problem? Any number of normative anti-discrimination theories focus on the reasons why members of protected groups are disadvantaged, asking questions like whether the discriminator was motivated by animus or other types of improper motivations.[120] Given that proxy discrimination by AIs is predominantly a risk when legally suspect factors are directly predictive of the discriminator's legitimate objectives, one might suggest that such discrimination is non-problematic under these theories.

This objection misses the mark because normative anti-discrimination theories that focus on discriminators' motives are a poor fit when it comes to laws that prohibit "rational discrimination."[121] In these cases, the law prohibits discrimination even though there is (arguably) nothing morally objectionable about the discriminator's logic for disfavoring members of the protected

119. *See* CLAUDE M. STEELE, WHISTLING VIVALDI AND OTHER CLUES TO HOW STEREOTYPES AFFECT US 125–26 (2010) (discussing how "stereotype threat" can "increase vigilance toward possible threat[s] and bad consequences in the social environment, which divert[] attention and mental capacity away from the task at hand").

120. *See, e.g.*, Larry Alexander, *What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies*, 141 U. PA. L. REV. 149, 175 (1992) (describing how "many otherwise immoral reaction preferences are preferences of individuals who are not fully morally responsible").

121. This point is detailed at length by Professor Samuel R. Bagenstos. *See* Bagenstos, *supra* note 85, at 836–37 (arguing that laws prohibiting "rational discrimination" cannot be coherently defended based on concerns regarding the discriminator's motivation, but must instead be justified based on outcome-oriented concerns such as mitigating "a pattern of social and economic subordination that has intolerable effects on our society").

group.[122] This point is particularly powerful when it comes to discrimination based on directly predictive characteristics.[123] When a discriminator relies on indirectly predictive characteristics (as in ordinary statistical discrimination), it uses group characteristics rather than exerting the effort to directly assess an individual's relevant traits. For that reason, the discriminator arguably engages in an unreasonable decision-making process that amounts to stereotyping.[124] But when an illicit characteristic is directly predictive, it is impossible for the discriminator to more directly assess the relevant characteristic, meaning that such group-based logic is hard to assail.[125] Accordingly, objections to treating individuals as members of groups have limited force in these settings.

For these reasons, the normative underpinnings of anti-discrimination regimes that prohibit discrimination based on directly predictive characteristics like disability, pregnancy, health, or genetics are necessarily predominantly outcome-oriented.[126] The goal of these laws, in other words, is to prevent socially-harmful outcomes for members of the protected group. It follows that proxy discrimination by AIs is normatively troubling because it will tend to produce the very results that the relevant anti-discrimination laws are designed to prevent.

The remainder of this Part details the various outcome-oriented reasons why the law might prohibit discrimination based on directly predictive traits. These include promoting social risk-sharing, preventing the chilling of socially valuable behavior, limiting the effects of past discrimination, and protecting non-conforming members of groups from being "actuarially saddled" with their group's characteristics. Although the relevance of these rationales varies across anti-discrimination regimes, the core point is that each is outcome-oriented, meaning that proxy discrimination by AIs will directly undermine the law's objectives. To illustrate, women who report experiencing intimate partner violence will find it harder to purchase life or property insurance; individuals with a pathogenic *BRCA* variant will face more limited insurance and employment prospects; individuals with disabilities will have a harder time securing employment; and minority students may find it harder

122. *See id.* Instead, the discriminator merely pursues the "ultimate end of maximizing profit [with] . . . no interest in harming minorities per se." *Id.* at 851.

123. *See, e.g.*, Mittelstadt et al., *supra* note 20, at 8 ("For the affected parties, data-driven discriminatory treatment is unlikely to be more palatable than discrimination fuelled [sic] by prejudices or anecdotal evidence. . . . [D]iscriminatory treatment is not ethically problematic in itself; rather, it is the effects of the treatment that determine its ethical acceptability.").

124. *See, e.g.*, Bagenstos, *supra* note 85, at 854–59.

125. *See* Peter J. Rubin, *Equal Rights, Special Rights, and the Nature of Antidiscrimination Law*, 97 MICH. L. REV. 564, 572–73 (1998) (proposing that discrimination laws make members of protected groups believe they are receiving special or equal treatment).

126. *See* Tal Z. Zarsky, *Understanding Discrimination in the Scored Society*, 89 WASH. L. REV. 1375, 1395 (2014) (arguing that the problem of blatant proxies is that they result in the very outcomes that original laws seek to prevent).

to secure admission to college. These outcomes, of course, strike at the heart of the underlying anti-discrimination laws to which they relate, irrespective of how or why they obtain.

1. Promoting Social Risk Sharing

A number of anti-discrimination regimes prohibit discrimination on the basis of directly predictive characteristics in order to socialize individual risks.[127] This goal is epitomized in the insurance context, where discrimination tends to undermine social risk sharing by fragmenting individuals into increasingly homogenous risk-pools.[128] Although such risk-based discrimination can help prevent moral hazard and adverse selection,[129] it can also impose undue or excessive risks on underserved groups.[130] Unregulated health insurance markets, for instance, typically result in those with substantial preexisting conditions being unable to acquire adequate

127. An extensive literature covers the goal of using anti-discrimination laws to achieve social solidarity by spreading certain risks, like the risk of negative health outcomes, across broad swaths of society. *See, e.g.*, Tom Baker, *Health Insurance, Risk, and Responsibility After the Patient Protection and Affordable Care Act*, 159 U. PA. L. REV. 1577, 1593–1602 (2011); Allison K. Hoffman, *Three Models of Health Insurance: The Conceptual Pluralism of the Patient Protection and Affordable Care Act*, 159 U. PA. L. REV. 1873, 1883–88 (2011). A closely related goal is promoting efficient redistribution through prohibitions on discrimination. *See* John Brooks, Brian Galle & Brendan Maher, *Cross-Subsidies: Government's Hidden Pocketbook*, 106 GEO. L.J. 1229, 1235–38 (2018); Kyle Logue & Ronen Avraham, *Redistributing Optimally: Of Tax Rules, Legal Rules, and Insurance*, 56 TAX L. REV. 157, 249 (2003); Ramsi Woodcock, Personalized Pricing and the Return of Wealth Redistribution at the Market Level 11–12 (2019) (unpublished manuscript), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3378864.

128. In the early history of insurance, carriers typically pooled the risks of community members without attempting to discriminate among them, thus converting insured risks from individual burdens into communal responsibilities. Tom Baker, *Containing the Promise of Insurance: Adverse Selection and Risk Classification*, 9 CONN. INS. L.J. 371, 372–73 (2003); Deborah A. Stone, *The Struggle for the Soul of Health Insurance*, 18 J. HEALTH POL. POL'Y & L. 287, 298–300 (1993); *see also* Austin, *supra* note 40, at 519–26. Such community insurance rating inevitably broke down in the face of competition, as new insurers sought to cherry-pick lower risk members of the community from the broader risk pool by offering them lower rates. *See generally* Peter Siegelman, *Adverse Selection in Insurance Markets: An Exaggerated Threat*, 113 YALE L.J. 1223 (2004) (arguing that adverse selection in insurance markets is not as great of a threat as is often predicted). This discrimination in favor of low-risk community members tended to become self-reinforcing; insurers who declined to discriminate among policyholders based on anticipated risks were left with increasingly high-risk policyholder, triggering increased premiums, and, ultimately, causing more relatively low-risk policyholders to be cherry-picked by competing insurers.

129. Adverse selection occurs when asymmetrical information allows high-risk individuals to enter an insurance pool at a premium level below their commensurate risk. Ronen Avraham, *The Economics of Insurance Law—A Primer*, 19 CONN. INS. L.J. 29, 44 (2012). Moral hazard is when policyholders take less care or do not minimize loss or risk of loss due to the fact that they have insurance to cover losses. *Id.* at 66.

130. *See, e.g.*, KENNETH S. ABRAHAM, DISTRIBUTING RISK: INSURANCE, LEGAL THEORY, AND PUBLIC POLICY 66 (1986) [hereinafter ABRAHAM, DISTRIBUTING RISK]; *see* Prince, *supra* note 26, at 631–32.

coverage at affordable prices.[131] Ensuring access to reasonable health insurance for these individuals not only protects some of the most vulnerable members of society, but also minimizes costs elsewhere in the system.[132]

It is for precisely these reasons that federal and state laws prohibit or limit discrimination by health insurers on the basis of numerous directly predictive characteristics.[133] By regulating discrimination on the basis of factors like preexisting conditions, age, and sex, state and federal laws seeks to achieve a specific outcome: the spreading of individual health risks across broad swaths of society so as to promote the availability of affordable health insurance.

Health insurance is hardly the sole example of a legal regime that prohibits discrimination on the basis of directly predictive traits in order to socialize risk. For instance, the ADA can also be justified on the basis that it properly shifts the costs of reasonable accommodations for individuals with disabilities to employers and society more broadly, so as to promote employment among those with disabilities.[134] Similarly, the goal of the PDA can largely be understood as partially socializing the employment-related costs of pregnancy, so that they are not borne entirely by women.

Proxy discrimination by AI strikes at the heart of regimes like these that seek to prohibit discrimination in order to promote social responsibility for certain risks. The reason should be obvious: They shift the costs of directly predictive characteristics back on to the protected group. Individuals with preexisting conditions may find it harder to purchase insurance; individuals with disabilities may be less able to secure employment; and women of child-bearing age may be paid less or have fewer employment opportunities. It is thus quite beside the point that proxy discrimination by AIs might produce these results without any conscious intent on the part of the discriminator or for reasons unrelated to animus or inaccurate stereotypes.

2. Preventing the Chilling of Socially Valuable Behavior

A second important reason why the law sometimes prohibits discrimination on the basis of directly predictive traits is to ensure that socially important activities are not chilled. This goal is most salient with respect to laws prohibiting discrimination on the basis of genetic information. As scientists first started mapping the human genome in the 1990s, advocates highlighted evidence showing that individuals were so fearful of genetic discrimination that they were avoiding genetic testing.[135] This fear, of course,

131. *See* Baker, *supra* note 128, at 377, 381.

132. *See* Emergency Medical Treatment and Active Labor Act, 42 U.S.C. § 1395dd (2012) (requiring covered hospitals to provide emergency medical care regardless of an individual's ability to pay).

133. *See supra* Section III.A.1.

134. *See* Bagenstos, *supra* note 85, at 839–44.

135. *See generally* Mark A. Hall, Jean E. McEwen, James C. Barton, Ann P. Walker, Edmund G. Howe, Jacob A. Reiss, Tara E. Power, Shellie D. Ellis, Diane C. Tucker, Barbara W. Harrison,

restricted the identification of important medical information that could be beneficial in research or clinical care.[136] It also impeded the acquisition of information that could help individuals take effective medical interventions to prevent or mitigate disease, avoid risky activities, and take drugs at doses that are particularly likely to be effective.[137] Congress passed GINA largely to counteract these concerns and encourage individuals to undertake genetic testing and participate in genetic research without fear of negative outcomes.[138]

GINA also helps to prevent the chilling of a different type of socially beneficial activity: the expressive or associational actions of those who learn that they have genetic risk factors. Such individuals are likely to rationally fear that their participation in potentially observable activities will trigger discrimination.[139] Those who have a pathogenic *BRCA1* or *BRCA2* variant, for instance, may choose to avoid looking for support communities because they legitimately fear that doing so may lead to future discrimination.[140]

Gordon D. McLaren, Andrea Ruggiero & Elizabeth J. Thomson, *Concerns in a Primary Care Population About Genetic Discrimination by Insurers*, 7 GENETICS MED. 311 (2005) (finding that concern about genetic discrimination varies substantially by race and other demographic factors and by nationality) [hereinafter Hall et al., *Concerns in a Primary Care Population*]; Mark A. Hall & Stephen S. Rich, *Patients' Fear of Genetic Discrimination by Health Insurers: The Impact of Legal Protections*, 2 GENETICS MED. 214 (2000) (finding that patients' and clinicians' fear of discrimination had not been limited by existing laws at the time of their survey); Yann Joly, Ida Ngueng Feze & Jacques Simard, *Genetic Discrimination and Life Insurance: A Systematic Review of the Evidence*, 11 BMC MED. 1 (2013) (finding fear of genetic discrimination prevalent in patients and research participants); E. Virginia Lapham, Chahira Kozma & Joan O. Weiss, *Genetic Discrimination: Perspectives of Consumers*, 274 SCI. 621 (1996) (finding a level of perceived discrimination in members of genetic support groups).

136. Areheart & Roberts, *supra* note 26, at 722.

137. *See, e.g.*, Allen D. Roses, *Pharmacogenetics and the Practice of Medicine*, NATURE, June 15, 2000, at 861 (describing the effectiveness of DNA-based screening).

138. Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233, 122 Stat. 881 (codified as amended in scattered sections of 26, 29, and 42 U.S.C) (noting potential avoidance of genetic testing as a reason for passing legislation); Areheart & Roberts, *supra* note 26, at 722–24. Of course, GINA can also be justified based on other goals, such as promoting social responsibility for genetically-encoded conditions. *See generally* Jessica L. Roberts, *Preempting Discrimination: Lessons From the Genetic Information Nondiscrimination Act*, 63 VAND. L. REV. 437 (2010) (describing some of the goals of GINA). For a broader discussion of the harms of genetic discrimination, see Susan M. Wolf, *Beyond "Genetic Discrimination": Toward the Broader Harm of Geneticism*, 23 J.L. MED. & ETHICS 345, 349–50 (1995).

139. *See* Joly et al., *supra* note 135, at 1–2; Hall et al., *Concerns in a Primary Care Population*, *supra* note 135, at 311; Laura M. Amendola, Jill O. Robinson, Ragan Hart, Sawona Biswas, Kaitlyn Lee, Barbara A. Bernhardt, Kelly East, Marian J. Gilmore, Tia L. Kauffman, Katie L. Lewis, Myra Roche, Sarah Scollon, Julia Wynn & Carrie Blout, *Why Patients Decline Genomic Sequencing Studies: Experiences from the CSER Consortium*, 27 J. GENETIC COUNSEL. ONLINE 1220, 1224 (2018).

140. As another example, a member of a particular political or religious group may avoid posting their group affiliation on social media or forego viewing a particular documentary or partaking in another action associated with the group out of fear of repercussions. Helveston, *supra* note 69, at 891–92.

GINA is not the only example of a law that prohibits discrimination on the basis of directly predictive characteristics so as to avoid chilling socially beneficial activities. Consider, for instance, state laws prohibiting insurers from discriminating on the basis of intimate partner violence.[141] A central explanation for such laws is that insurance discrimination could dissuade victims of violence from seeking needed medical care or police intervention.[142]

Proxy discrimination by AIs holds the potential to undermine these goals by allowing discriminators to indirectly harvest the predictive power of suspect traits like genetic tests or domestic violence reports. Any number of data points might allow an AI to proxy for such information, including the websites an individual visits, the location and information in their cell phones, or their social media posts.[143] Once individuals learned from experience or news reports, or even began to suspect, that activities like genetic testing or reporting domestic violence could result in future discrimination, proxy discrimination by AIs would tend to produce the very same results that the law sought to avoid: Individuals would decline to participate in socially-beneficial activities like genetic testing because they rationally fear the negative results that may follow.

To be sure, the ultimate impact of proxy discrimination by AIs on behavior is hard to fully anticipate. On one hand, the black box nature of AIs may minimize any particular chilling effect. In most instances of intentional discrimination or implicit bias, members of protected groups have an opportunity to understand the link between their protected status and an adverse event. By contrast, the link between a negative outcome and the specific data relied on by an AI is typically completely opaque to impacted individuals. This is for a variety of reasons, most notably the proprietary nature of most AIs and the vastness of the data on which they rely.[144] The upshot of this opacity is that many members of protected groups may not know enough about how or when AIs will attempt to proxy for their protected traits to adjust their behavior accordingly.

On the other hand, the opacity of AI and big data could plausibly produce much stronger chilling effects for members of protected groups than a more transparent system of discrimination. Those who experience anxiety about "revealing" their status to an AI could well adjust their behavior even more than necessary to avoid such discrimination. This is particularly likely when individuals have an intuitive understanding that their membership in a protected class could indeed be highly relevant to firms' facially neutral goals.

---

141. *See generally* Hellman, *supra* note 96 (examining the claims from the insurance industry and its critics regarding insurance for battered women).

142. *Id.* at 376–77.

143. *See* David C. Vladeck, *Consumer Protection in an Era Of Big Data Analytics*, 42 OHIO N.U. L. REV. 493, 497–501 (2016) (describing how data brokers collect and store information).

144. *See supra* Part II.

Thus, cancer survivors or individuals with genetic markers for Huntington's disease may be particularly likely to refrain from activities associated with these facts, anticipating the mere possibility that an insurer, credit institution, or employer could harvest information on those activities.

These harms are not just theoretical. Recently, life insurers have started predicting life expectancy by relying on proxies that derive from social media.[145] This reality has led prominent newspapers like the Wall Street Journal to recommend that individuals post on social media pictures of themselves exercising and eating healthy, while avoiding posts of themselves smoking or engaging in extreme sports.[146] As proxy discrimination by AI becomes more common, it is easy to imagine similar newspaper stories warning individuals not to join Facebook groups associated with suspect characteristics like genetic conditions or domestic violence, because doing so might result in future adverse consequences for insurance, credit, or employment.[147]

3. Limiting or Reversing the Effects of Past Discrimination

Another reason why the law may forbid discrimination based on directly predictive characteristics is to slow an otherwise self-replicating pattern of economic subordination experienced by members of historically disadvantaged groups.[148]

Anti-subordination goals are particularly relevant with respect to prohibitions on the use of race, even when race is directly predictive of legitimate considerations, like recidivism rates or predicted academic

---

145. *See* Leslie Scism, *New York Insurers Can Evaluate Your Social Media Use—If They Can Prove Why It's Needed*, WALL ST. J. (Jan. 30, 2019, 9:00 AM), https://www.wsj.com/articles/new-york-insurers-can-evaluate-your-social-media-useif-they-can-prove-why-its-needed-11548856802 [https://perma.cc/A9RK-QEH3].

146. *See id.*

147. Prominent politicians have also warned of the discriminatory harms of big data and AI. *See* Danny Li, *AOC Is Right: Algorithms Will Always Be Biased As Long As There's Systemic Racism in This Country*, SLATE (Feb. 1, 2019, 3:47 PM), https://slate.com/news-and-politics/2019/02/aoc-algorithms-racist-bias.html [https://perma.cc/MS6F-YYGM].

148. *See* Owen M. Fiss, *A Theory of Fair Employment Laws*, 38 U. CHI. L. REV. 235, 260 (1971) (discussing employer's liability in the antidiscrimination context); Sunstein, *supra* note 29, at 8. Of course, a vast literature exists exploring this antisubordination view of antidiscrimination law. *See, e.g.*, Robert Post, *Prejudicial Appearances: The Logic of American Antidiscrimination Law*, 88 CAL. L. REV. 1, 30–31 (2000) (proposing an understanding of antidiscrimination law premised on changing social practices). *See generally* Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over* Brown, 117 HARV. L. REV. 1470 (2004) (revisiting the role antisubordination and antidiscrimination values play in the post-*Brown* equal protection framework). So too does a literature that rejects this view of antidiscrimination law, favoring instead an autoclassification logic that focuses on prohibiting decision-making based on impermissible factors. *See, e.g.*, Kelman, *supra* note 102, at 845–46. As suggested by the earlier discussion, whatever the merits of this debate in general, this type of anti-classification logic is a poor fit when it comes to laws that prohibit discrimination on the basis of traits that are directly predictive of otherwise legitimate goals of the discriminator. *See supra* Section III.B.

performance. For instance, reconsider research demonstrating that members of certain minority groups tend to experience self-fulfilling anxiety that their academic performance will confirm negative prejudices.[149] As discussed above, this phenomenon potentially causes race to be directly predictive of college performance.[150] At the same time, negative prejudices about racial groups are themselves a product of historical animus and subordination. By forbidding discrimination based on race, even though it is in fact directly predictive of otherwise legitimate factors like anticipated college performance, the legal system attempts to limit the capacity of past discrimination to impact future results.

This goal of limiting the impact of historical subordination animates prohibitions on directly predictive forms of discrimination in other domains as well. For instance, federal prohibitions on pregnancy discrimination were generally justified as necessary to overcome workplace structures that were designed by men for men.[151] Similar arguments have often been made to justify the ADA, as many of the difficulties that individuals with disabilities face in traditional work environments are themselves a legacy of those with disabilities being excluded from traditional employment settings.[152]

As above, proxy discrimination by AIs undermines this anti-subordination goal by precluding the realization of the law's objectives.[153] Laws that are based on anti-subordination principles are fundamentally about changing social and economic structures that reflect and reinforce historical discrimination. Proxy discrimination by AIs affirmatively thwarts this objective by reproducing and reinforcing these legacies of historical discrimination on the implicit ground that they make economic sense for discriminators. Yet the rationale of these laws prohibiting discrimination on the basis of directly predictive traits is that change is often necessary, even though it can be costly or difficult for those who benefit most from the existing system.

Not only can proxy discrimination by AIs thwart the anti-subordination goals of existing anti-discrimination laws, but it can affirmatively promote the opposite result. By allowing discriminators to indirectly but reliably take into account the ways in which historical discrimination impacts marginalized groups, proxy discrimination by AIs can cloak the reproduction of these

149. *See supra* text accompanying note 119.

150. We do not imply, of course, that race is causally predictive of educational performance. However, it may be directly predictive through opaque relationships given that the impact of past discrimination and societal structures is difficult to quantify and measure.

151. *See* Siegel, *supra* note 104, at 951–52.

152. *See* Bagenstos, *supra* note 85, at 839.

153. *See, e.g.*, Sunstein, *supra* note 29, at 8 ("Difficult problems are also presented if an algorithm uses a factor that is in some sense an outgrowth of discrimination. For example, a poor credit rating, or a troubling arrest record, might be an artifact of discrimination, by human beings, before the algorithm was asked to do its predictive work. There is a risk here that algorithms might perpetuate discrimination, and extend its reach, by using factors that are genuinely predictive, but that are products of unequal treatment.").

historical hierarchies in seemingly neutral and objective structures. For instance, minority job applicants may face difficulty beating employment algorithms that proxy discriminate for race due to the reality that past minority applicants may have faced difficult adjustment periods due to factors like stereotype threat.[154] This lack of steady employment can lead to limited credit availability, causing difficulties getting insurance and access to healthcare. And these realities, in turn, can cycle back to making college even less accessible to targeted members of historically-disadvantaged groups.[155] This type of feedback loop makes proxy discrimination by AIs particularly pernicious, since it is the inequitable outcome from one silo that makes the use of that outcome as a proxy rational in the next silo.

4. Anti-Stereotyping

Another potential goal of anti-discrimination regimes that prohibit the use of directly predictive characteristics is to prevent the classification of individuals based on their membership in certain stereotyped groups. As suggested above, such an anti-stereotyping principle is hard to justify based on the impropriety of the discriminator's decision-making process when the suspect characteristic is directly predictive.[156] But anti-stereotyping can be a coherent goal of anti-discrimination regimes that prohibit use of directly predictive traits to the extent that the focus is on the potential unfairness of the outcomes produced by such stereotyping. Even rational discrimination based on directly predictive traits necessarily results in individuals who do not conform to group averages being treated as if they do.[157]

Numerous court cases highlight this tension between the rational use of averages in models and the desire for individualized treatment. For example, in *City of Los Angeles, Department of Water and Power v. Manhart*, the Supreme Court reviewed a pension system where female employees paid larger contributions than men for the same monthly benefit due to higher life expectancies.[158] The majority ultimately determined that this scheme violated

154. *See* Claude M. Steele & Joshua Aronson, *Stereotype Threat and the Intellectual Test Performance of African Americans*, 69 J. PERSONALITY & SOC. PSYCHOL. 797, 810 (1995) (concluding "that stereotype threat is an underappreciated source of classic deficits in standardized test performance").

155. *See* O'NEIL, *supra* note 5, at 147–49.

156. *See supra* text accompanying notes 121–26. An anti-stereotyping principle can also be justified by reducing social stigma for members of a protected group. However, given the opacity of AI, this stigmatization may be a less-likely harm of proxy discrimination than other concerns of anti-stereotyping.

157. *See* ABRAHAM, DISTRIBUTING RISK, *supra* note 130, at 74–75; *see also* Bornstein, *supra* note 7, at 525–28 (arguing that the anti-stereotyping theory of Title VII could be used to limit some harms of algorithmic decision-making in employment).

158. City of L.A. Dep't of Water & Power v. Manhart, 435 U.S. 702, 704–05, 708 (1978) ("The question, therefore, is whether the existence or nonexistence of 'discrimination' is to be determined by comparison of class characteristics or individual characteristics.").

Title VII of the 1964 Civil Rights Act, because it assumed individuals would conform to broader trends associated with their sex.[159] Such discrimination, the court suggested, is troubling from a civil rights perspective because it fails to treat individuals as individuals, as opposed to merely members of the groups to which they belong.

Of course, the contexts in which the law tolerates the potential unfairness of attributing group characteristics to individuals varies across contexts and groups. As *Manhart* suggests, employers are forbidden from stereotyping based on a wide range of characteristics, including: age, disability, race, sex, and genetic information.[160] These laws are driven, in part, by the fact that employment decisions are generally individual: A specific person is hired, fired, or demoted, based on his or her past or expected contribution to the employer's mission. By contrast, stereotyping individuals based on group characteristics is generally more tolerated in domains like insurance, where individualized decision-making is often impractical.[161]

As discussed in Part III, proxy discrimination by AIs can directly undermine the law's efforts to limit the unfair outcomes of stereotyping for non-conforming members of the group. In some cases, AI could minimize stereotype harm if more predictive variables are available. However, in other cases, especially when the predictive power of the stereotype is opaque or direct, algorithms directly target members of protected groups and then assign them the characteristics of that group. In such cases, proxy discrimination by AIs "actuarially saddles" members of a protected group with the general characteristics of their group.[162]

---

159. *Id.* at 708–11. In his concurrence, Justice Blackmun voiced his discomfort with this rationale, arguing that an individualized analysis is unrealistic because there is no way to accurately predict when someone will die. *Id.* at 724 (Blackmun, J., concurring). Similarly, Justice Burger's dissent noted that since it is impossible to make individual determinations about lifespan, the use of actuarial data is an attempt "to treat them as individually as it is possible to do in the face of the unknowable length of each individual life." *Id.* at 727–28 (Burger, J., dissenting).

160. *See* Bornstein, *supra* note 7, at 525–26 (arguing that predictive AIs in the employment setting can be challenged under an anti-stereotyping theory of disparate treatment law); Jessica A. Clarke, *Beyond Equality? Against the Universal Turn in Workplace Protections*, 86 IND. L.J. 1219, 1225 (2011); Kim, *supra* note 3, at 884–85;

161. *See generally* Avraham et al., *supra* note 38 (arguing that while there are limits to stereotyping by insurance companies, stereotyping is how different risk groups are identified).

162. *Hartford Accident & Indemnity Co. v. Insurance Commissioner of Pennsylvania* is illustrative of the potential unfairness of this approach. Hartford Accident & Indem. Co. v. Ins. Comm'r of Pa., 482 A.2d 542, 545–46 (Pa. 1984). In that case, male auto policy holders complained that they were charged higher premiums for the same coverage as women of the same age and driving records. *Id.* The Commissioner found that such gender-based premiums in auto insurance constituted "unfair discrimination," and the court agreed, in part because there is a lack of causality between gender and accidents. *Id.* Statistical calculations do exactly this, because they only consider the likelihood that, on average, individuals with a specified trait will experience the outcome in question. *See* Abraham, *Efficiency and Fairness*, *supra* note 24, at 408; *see also* FREDERICK SCHAUER, PROFILES, PROBABILITIES, AND STEREOTYPES 3–7 (2006); Barocas & Selbst, *Big Data*,

* * *

The Figure below summarizes the analysis in this Part demonstrating the implications of proxy discrimination by AIs.

Figure 3

Prohibition on use of characteristic that may be "directly predictive" of legitimate goal.	Outcome oriented rationales for legal prohibition on use of "directly predictive" traits.	Example of how proxy discrimination by AI can undermine goal of underlying anti-discrimination regime.
GINA and state law prohibitions on genetic information	(1) Encouraging socially valuable behavior; (2) Promoting social risk sharing	AIs used by employers proxy for genetic information, causing decreased genetic testing.
State insurance prohibitions on domestic violence history, sex, and race	(1) Encouraging socially valuable behavior; (2) Limiting effects of past discrimination; (3) Anti-Stereotyping	AIs used by insurers proxy for domestic violence history, causing victims of such violence to pay higher rates.
State and federal prohibitions on use of medical history, sex in health insurance	(1) Promoting social risk sharing	Health insurer AI proxies for preexisting condition in decreasing marketing in areas, making coverage less available for high risk.
ADA prohibition on discrimination against disabled	(1) Promoting social risk sharing; (2) Limiting effects of past discrimination	Employers use AIs that proxy for disability, causing individuals with disabilities to have harder time finding employment.
APA prohibition on discrimination against pregnant women	(1) Promoting social risk sharing; (2) Limiting effects of past discrimination; (3) Anti-Stereotyping	Employers use AIs that proxy for pregnancy, causing pregnant women and women of child-bearing age to have harder time finding employment.
ADEA prohibition on discrimination based on age	(1) Promoting social risk sharing; (2) Anti-Stereotyping	Employers use AIs that proxy for age, causing older workers to have harder time finding employment.
State and federal prohibitions on use of race in context of stereotype threat.	(1) Limiting effects of past discrimination; (2) Anti-Stereotyping	University uses AI that proxies for race given reality of stereotype threat, making higher education less available to African Americans.

supra note 5, at 688 ("As Professor Frederick Schauer explains, decision makers that rely on statistically sound but non-universal generalizations 'are being simultaneously rational and unfair' because certain individuals are 'actuarially saddled' by statistically sound inferences that are nevertheless inaccurate."); Cary Franklin, *The Anti-Stereotyping Principle in Constitutional Sex Discrimination Law*, 85 N.Y.U. L. REV. 83, 120 (2010) ("[Legal feminists] needed an approach that would direct courts' attention to the particular institutions and social practices that had perpetuated inequality in the context of sex and counteract the widespread perception that sex discrimination redounded to women's benefit."); Cary Franklin, *Inventing the "Traditional Concept" of Sex Discrimination*, 125 HARV. L. REV. 1307, 1354–58 (2012) (discussing interpretations of Title VII as a means of combatting gender-based discrimination in the workplace).

IV. RESPONDING EFFECTIVELY TO PROXY DISCRIMINATION

As Parts II and III make clear, the accelerating evolution of AI and big data render proxy discrimination a fundamental threat to important goals of many, if not most, antidiscrimination regimes. As such, this Part considers a variety of potential options for how antidiscrimination regimes might respond to the emerging risk of proxy discrimination by AIs. Section IV.A begins by explaining why two common features of antidiscrimination regimes—a ban on the use of obvious proxies for suspect characteristics and disparate impact liability—cannot effectively prevent proxy discrimination by AI. Section IV.B then surveys five more promising approaches for combatting the risk of proxy discrimination by AIs. These strategies either impact the data that AIs can access or regulate when or how AIs can use this data.[163]

A. INEFFECTIVE SOLUTIONS

Many antidiscrimination regimes have features that are capable of policing against traditional, intentional proxy discrimination against protected groups, such as red lining. The two most pervasive such strategies are to explicitly ban the use of specific potential proxies and to subject discriminators to a disparate impact theory of liability. As we describe below, however, neither of these strategies has any plausible chance of combatting proxy discrimination by AIs.[164]

1. Ban Discriminators' Use of Obvious Proxies for Protected Characteristics

Many antidiscrimination regimes ban actors not just from utilizing a protected trait, but also from considering obvious proxies for this protected trait. GINA exemplifies this strategy. In GINA, Congress recognized that simply banning insurer use of genetic test results would do little to assuage public fear of discrimination if employers and insurers could substitute clear proxies for genetic results into their decisions. For example, a law that prohibits employers from discriminating on the basis of a test result indicating increased risk of colon cancer does little if employers could simply extrapolate the likely genetic status of the individual from family history. GINA, therefore, not only bars the use of genetic test results, but also the use of several of the most obvious proxies for this information. GINA accomplishes this by broadly

163. This menu of options explores only the narrow concerns of algorithmic proxy discrimination. There are a host of other potential concerns with bias, skewed data, and discriminatory impacts of algorithms at large. Regulatory options should consider and address these broader concerns, but this Article focuses on solutions that may address the specific concerns of algorithmic proxy discrimination, some of which may help address broader concerns as well. *See* Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. (forthcoming 2019).

164. Because we assume that algorithms cannot intentionally discriminate, we do not discuss legal prohibitions on such discrimination in this Part. *See supra* note 15 and accompanying text.

defining "genetic information" to include a spectrum of genetic-related traits, such as genetic test results, family medical history, participation in genetic research, and use of genetic services, such as going to see a genetic counselor.[165]

GINA not only forbids employers and health insurers from using any genetic information, but also limits them from collecting this information.[166] These privacy protections make GINA distinct among most antidiscrimination laws in the employment setting, where information about protected traits is readily observable to discriminators.[167] By restricting the availability of protected information and obvious proxies for that information, GINA attempts to limit the capacity of employers, insurers, or other actors to discriminate against protected individuals.[168]

State insurance law also attempts to combat proxy discrimination by banning insurers' consideration of obvious proxies for prohibited characteristics, as well as their access to information about those characteristics. The exact contours of this strategy vary by state and line of coverage. Most states ban insurers from collecting any information about suspect characteristics, like race or income.[169] Additionally, as with GINA, some states ban insurers from using specific proxies for protected characteristics. Prohibitions on insurer consideration of credit score (arguably a potential proxy for policyholder race/income) and zip code (a more concerning proxy for policyholder race/income) are illustrative.[170] Finally, state regulators and policymakers occasionally scrutinize known classification factors that could be proxies for suspect characteristics.[171] For

165. Genetic Information Nondiscrimination Act, 42 U.S.C. § 2000ff(4) (2012).

166. *Id.* § 2000ff-1(b).

167. *See generally* Jessica L. Roberts, *The Genetic Information Nondiscrimination Act as an Antidiscrimination Law*, 86 NOTRE DAME L. REV. 597 (2011) (evaluating GINA as an antidiscrimination law).

168. This strategy is supported empirically as a way to address issues of bias, whether direct or implicit. *See generally, e.g.*, Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991 (2004) (showing that job applicants with African-American sounding names on resumes were less likely to be interviewed than those with White-sounding names, even though the resume qualifications were similar); Claudia Goldin & Cecilia Rouse, *Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians*, 90 AM. ECON. REV. 715 (2000) (showing that more female musicians were selected for orchestras when their gender was hidden from view).

169. *See* Daniel Schwarcz, *Towards a Civil Rights Approach to Insurance Anti-Discrimination Law*, 69 DEPAUL L. REV. (forthcoming 2020) (manuscript at 11–12, 24–25) [hereinafter Schwarcz, *Civil Rights*].

170. *See* Austin, *supra* note 40, at 525–26: Squires, *supra* note 37, at 392; Works, *supra* note 40, at 472.

171. Thus, a regulatory handbook for insurance examiners instructs them to identify "any 'red flags,' such as . . . a factor that is an obvious proxy for some prohibited characteristic" when reviewing insurers' underwriting and rating practices. NAT'L ASS'N OF INS. COMM'RS, MARKET REGULATION HANDBOOK 63 (2017), *available at* https://www.in.gov/idoi/files/Market%20Regulation%20Handbook%2017_Vol1.pdf [https://perma.cc/G6SJ-Z5GV]. When such red

instance, insurers' reliance on credit information became controversial only after it was recognized that insurers' use of this information might operate as a proxy for legally-suspect characteristics, like income or race.[172]

Prohibitions on discriminators' consideration of potential proxies for suspect characteristics also appear outside the insurance setting. For example, two acts recently passed in New York[173] and California seek to expand the definition of race to include hair texture and hairstyles because these are traits "historically associated with race."[174] As the findings of the legislation state, "[i]n a society in which hair has historically been one of many determining factors of a person's race, and whether they were a second class citizen, hair today remains a proxy for race."[175] For this reason, the laws bar employers and educational institutions from discriminating against those with hairstyles common to African-Americans, such as braids, locks, and twists.

Although these strategies may effectively prevent traditional intentional proxy discrimination,[176] they have little power to prevent proxy discrimination

---

flags exist, regulators are supposed to ask whether "the underwriting guideline serve[s] a necessary underwriting purpose by identifying a characteristic of the consumer, vehicle or property that is demonstrably related to risk of loss and does not duplicate some other factor that has already been taken into account." *Id.*

172. BIRNY BIRNBAUM, CTR. FOR ECON. JUSTICE, INSURANCE CREDIT SCORING: AN UNFAIR PRACTICE 6 (2005), *available at* http://www.cej-online.org/cej%20report%20ins%20cr%20scoring%200501.pdf [https://perma.cc/9BEF-HMA9]; *see, e.g.*, FED. TRADE COMM'N, CREDIT-BASED INSURANCE SCORES: IMPACTS ON CONSUMERS OF AUTOMOBILE INSURANCE 51–56 (2007), *available at* https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf [https://perma.cc/BB9Q-668G] (analyzing the relationship between credit scores and race and income).

173. Press Release, Governor Andrew M. Cuomo, Governor Cuomo Signs S6209A/A7797A to Make Clear Civil Rights Laws Ban Discrimination Against Hair Styles or Textures Associated with Race (July 12, 2019), *available at* https://www.governor.ny.gov/news/governor-cuomo-signs-s6209aa7797a-make-clear-civil-rights-laws-ban-discrimination-against-hair [https://perma.cc/ZR9H-FYXS].

174. *See* D. Wendy Greene, *Title VII: What's Hair (and Other Race-Based Characteristics) Got to Do With It?*, 79 U. COLO. L. REV. 1355, 1385 (2008) (describing how an employer's prohibition of certain hairstyles associated with blackness can "demonstrate a prima facie case of race discrimination"). *See generally* CAL. EDUC. CODE § 212.1 (2020) (amending the California code to preclude such discrimination).

175. Crown Act, S.B. 188 § 1(f) (Cal. 2019) (amending CAL. EDUC. CODE § 212.1).

176. Indeed, we are by no means arguing that GINA or the Crown Acts are futile overall. As a method to combat disparate treatment and intentional proxy discrimination by humans, it may achieve its goal. For example, although GINA was heralded as an important civil rights bill, its success at protecting against discrimination on the basis of genetic test results has been very limited. Areheart & Roberts, *supra* note 26, at 725, 730. Indeed, a review of case law in the first ten years of the law indicated no claims of employment adverse events on the basis of genetic test results. *Id.* at 730. Instead, plaintiffs to date have argued that employers discriminated on the basis of family history or that employers violated the privacy provisions of the law. *Id.* at 735–36, 755. Thus, a predominant part of the GINA caselaw has focused on employer collection and use of proxy variables for genetic information, leaving the question of whether GINA was needed and whether genetic test results are ever being used by employers. *Id.* at 750–51. In this way, GINA

by AIs. As suggested in Part II, AIs that are deprived of direct information about suspect characteristics and obvious proxies for this information will inevitably identify other proxy variables for directly predictive data.[177] Simply removing an additional set of obvious proxies only forces the algorithm to find slightly less intuitive or slightly less accurate proxies in their stead.[178] This point is well illustrated by the fact that the contributions made by individual variables in AI models routinely change depending on the training data on which they rely.[179] For example, an AI that does not have access to data on positive genetic test results or visits to a genetic counselor could just as easily rely on membership in a genetics-community social media page to proxy for this directly predictive information. Similarly, to the extent that race was directly predictive of a target variable, an AI that did not have access to information about race or hairstyles would inevitably tend to construct

has arguably been somewhat successful at addressing proxy discrimination as it has been historically conceptualized—where a particular actor specifically chooses to employ a proxy for a trait he or she can no longer consider.

177. *See supra* Section II.C.

178. In recent years, state insurance regulators have occasionally acknowledged this substantial gap in their regulatory scheme. For instance, the newly appointed NAIC President recently opined, "We want to encourage innovation but can't allow models to be proxies for things which could be discriminatory practices." National Association of Insurance Commissioners (@naic), TWITTER (Jan. 11, 2019, 8:34 AM), https://twitter.com/naic/status/1083764029485731840 [https://perma.cc/V25B-P5H5]. Towards that end, the NAIC developed a "Big Data" working group that is in the process of developing a white paper on best practices for "[r]egulatory [r]eview of [p]redictive [m]odels." *See Regulatory Review of Predictive Models* 1 (National Association of Insurance Commissioners, 2019), *available at* https://content.naic.org/sites/default/files/inline-files/Predictive%20Model%20White%20Paper%20Exposed%208-3-19.pdf [https://perma.cc/UVK4-XETX]. Remarkably, though, the current white paper draft does not identify the unique risks of algorithmic proxy discrimination, whether or how state regulators should attempt to identify it, or whether it may violate state laws. Instead, it simply directs state regulators to consider whether any input or output data is "unfairly discriminatory," a requirement that—unadorned without further comment—requires only that there exist an actuarial relationship between the input data and claims projections. *Id.* at 5. The lack of any coherent framework for identifying, diagnosing, or responding to proxy discrimination in insurance is particularly troubling because insurance markets are likely to aggressively exploit AI and big data to discriminate among policyholders. *See* Ari Libarikian, Kia Javanmardian, Doug McElhaney & Ani Majumder, *Harnessing the Potential of Data in Insurance*, MCKINSEY & CO. (May 2017), https://www.mckinsey.com/industries/financial-services/our-insights/harnessing-the-potential-of-data-in-insurance [https://perma.cc/D4F9-WHDD]; Dan Robinson, *AI in Insurance: How Artificial Intelligence and Big Data Could Transform Sector in 2019*, NS BUS. (Dec. 14, 2018), https://www.ns-businesshub.com/science/ai-in-insurance-2019 [https://perma.cc/SW73-6Q5A]; *see also* Michael W. Elliott, *Insights 2017 Article: Big Data Analytics: Changing the Calculus of Insurance*, INSTS.: CPCU SOC'Y (June 23, 2017, 10:29), https://infotech.ig.cpcusociety.org/news/insights-2017-article-big-data-analytics-changing-calculus-insurance [https://perma.cc/S2LZ-9EDY]. For these reasons, the Government Accountability Office recently highlighted proxy discrimination as a potential concern in the growing insuretech sector. U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-19-423, INSURANCE MARKETS: BENEFITS AND CHALLENGES PRESENTED BY INNOVATIVE USES OF TECHNOLOGY 17 (2019).

179. *See* Gillis & Spiess, *supra* note 29, at 463.

alternative proxies, such as Netflix shows watched, or even hair products purchased.

Making matters even worse, the black box nature of AIs and the vastness of big data mean that intuition alone will often be inadequate to identify an AI's use of a proxy variable, even after the fact.[180] No longer are the "traditional" proxies, like headgear, hairstyles, or height and weight, the only potential substitutes for our society's protected traits. Instead, AIs can generate proxies for directly predictive suspect traits based on all sorts of behavior, from what movies one streams online to the language one uses in social media posts. Even more importantly, the proxies available to AIs may consist of numerous interacting pieces of data, whose significance as a proxy may be completely unintuitive.[181] For instance, people with a pathogenic variant in the *BRCA1* or *BRCA2* genes may be identifiable to an AI that combines geo-locational, web-surfing, and shopping patterns.[182] For this reason it will often be impossible to determine whether an AI is proxying for a protected trait simply by scrutinizing the data on which it ultimately relies.[183]

2. Traditional Disparate Impact Liability

A second common strategy for combatting proxy discrimination is disparate impact liability.[184] Of course, such liability is particularly important in the employment context, where Title VII bars employment practices that have a disparate impact based on race, color, religion, national origin, or sex.[185] But disparate impact liability also exists in a number of other anti-discrimination regimes, including housing and credit.[186] By contrast,

180. This point is true even though, as Gillis & Spiess emphasize, the decision rule is actually much more transparent in the context of discrimination by AIs, as compared to discrimination by humans. *See id.* at 465.

181. W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419, 421 (2015) ("[M]any important relationships are not one-to-one, two-to-one, or even several-to-one correspondences, but are instead networks among dozens of interacting variables, including those which are readily observable . . . and those that are not . . . ."); *see also* Coglianese & Lehr, *supra* note 59, at 17 ("[C]omplex data sets necessarily contain complex inter-variable relationships, making it even more difficult to put into intuitive prose how a machine-learning algorithm makes the predictions it does.").

182. SOLON BAROCAS, ALEX ROSENBLAT, DANAH BOYD, SEETA PEÑA GANGADHARAN & CORRINE YU, DATA & CIVIL RIGHTS: TECHNOLOGY PRIMER 1 (Oct. 30, 2014), http://www.datacivilrights.org/pubs/2014-1030/Technology.pdf [https://perma.cc/5ABC-N6T6]; Coglianese & Lehr, *supra* note 59, at 17; Price, *supra* note 181, at 42.

183. Glymour & Herington, *supra* note 79, at 275.

184. For a description of the basic disparate impact legal framework, see *supra* Part II.

185. *See supra* Section III.A.3.

186. *See* Michael Aleo & Pablo Svirsky, *Foreclosure Fallout: The Banking Industry's Attack on Disparate Impact Race Discrimination Claims Under the Fair Housing Act and the Equal Credit Opportunity Act*, 18 B.U. PUB. INT. L.J. 1, 22–38 (2008) (discussing disparate impact theory in housing and "lending discrimination cases under the FHA and ECOA").

disparate impact liability is not available under state insurance laws or GINA.[187]

Disparate impact liability does indeed help combat intentional proxy discrimination. This should not be surprising, given that proxy discrimination is simply one specific type of practice that produces a disparate impact.[188] To be sure, discriminators can defeat a disparate impact claim by showing that their practices are consistent with business necessity and that no less discriminatory alternative is available.[189] But meeting these burdens will typically be difficult for an intentional proxy discriminator, especially since the plaintiff can show that any such explanation is pretextual. At least in part for these reasons, a number of commentators have even suggested that the core goal of disparate impact regimes is to help identify shrouded intentional discrimination, a category that includes intentional proxy discrimination.[190]

By contrast, disparate impact liability (as it is currently constructed) is simply not capable of effectively policing against proxy discrimination by AIs. The central problem is that firms using AIs that proxy discriminate will typically have little problem showing that this practice is consistent with business necessity and in rebuffing any attempt to show the availability of a less discriminatory alternative.[191] This is because, by definition, proxy discrimination helps the AI predict a legitimate objective: the target variable it is programmed to optimize, like anticipated insurance claims.[192] Moreover, there is no obvious way for a plaintiff to advance a less discriminatory alternative, given that AIs are indeed uniquely effective at optimizing their programmed objective, notwithstanding their tendency to construct proxies

187. GINA expressly excludes a private cause of action based on disparate impact. 42 U.S.C. § 2000ff-7(a) (2012); Ifeoma Ajunwa, *Genetic Data and Civil Rights*, 51 HARV. C.R.-C.L. L. REV. 75, 86–87 n.74 (2016) ("As of the writing of this Article, Congress has yet to establish the commission as mandated by GINA."); Jennifer K. Wagner, *Disparate Impacts and GINA: Congress's Unfinished Business*, 5 J.L. & BIOSCI. 527, 545 (2019); Schwarcz, *Civil Rights*, *supra* note 169, at 19. In insurance, disparate impact liability has historically potentially been available against property insurers under the Fair Housing Act. *See* Kaersvang, *supra* note 8, at 1997. But such liability has also faced a number of important hurdles, including reverse preemption under the McCarran Ferguson Act.

188. *See supra* Part II (describing how proxy discrimination is one particular, and unusually pernicious, form of disparate impact).

189. Civil Rights Act ("Title VII") of 1964, 42 U.S.C. § 2000e-2(k)(1)(A); Girardeau A. Spann, *Disparate Impact*, 98 GEO. L.J. 1133, 1149 (2010).

190. *See supra* Part II.

191. *See* Susan S. Grover, *The Business Necessity Defense in Disparate Impact Discrimination Cases*, 30 GA. L. REV. 387, 392 (1996) (describing how the Civil Rights Act of 1991 shapes the business necessity defense in disparate impact cases).

192. Kim, *supra* note 3, at 866 (noting that "to ask whether the model is 'job related' in the sense of 'statistically correlated' is tautological"); *see* Bornstein, *supra* note 7, at 553–58; *see also* Barocas & Selbst, *Big Data*, *supra* note 5, at 701–12 (discussing the potential for discrimination in data mining and AI).

for directly predictive suspect characteristics to which they do not have access.[193]

B. Potentially Effective Strategies for Combatting Proxy Discrimination by AIs

Although traditional approaches to combatting intentional proxy discrimination are inadequate to prevent proxy discrimination by AIs, the law is not powerless to prevent such discrimination. Below we discuss five different strategies that might be able to effectively combat proxy discrimination by AIs, depending on the context. The first two approaches are mutually exclusive as they relate to the amount and type of data actors can access. By contrast, the next three possibilities—which require algorithms to be transparent, ethical, or justified by plausible causal connections—condition firms' use of AIs to discriminate in a manner that could be combined with one another, as well as coupled with one of the first two options.

Of course, no single solution will be appropriate for all anti-discrimination regimes. Instead, the optimal interventions will depend on myriad factors, such as the extent to which proxy discrimination is likely to strike at the heart of a particular anti-discrimination regime's goals and the existing infrastructure for policing against prohibited forms of discrimination. Additionally, because the goal of algorithms is to ferret out the most efficient predictors of a programmed outcome, any regulatory interventions will naturally limit discriminators' capacity to achieve their otherwise legitimate goals. However, algorithmic proxy discrimination can only exist when the law has decided to prohibit "rational" discrimination due to broader social concerns.[194] If algorithmic proxy discrimination is left unchecked due to narrowly-defined notions of efficiency, then it must be acknowledged that this comes at the expense of these laws' goals.

1. Flipping the Default: Prohibiting Discrimination Based on Non-Approved Factors

As suggested above, in an age of AI and big data it is impossible to identify ex ante all potential proxies for suspect characteristics, as GINA and other laws attempt to do. Proxy discrimination by AIs could thus be prevented by flipping the default approach of anti-discrimination law: Instead of allowing use of any variable not barred, as in the traditional anti-discrimination model, this approach would only allow actors to use pre-approved variables. It would thus limit algorithmic proxy discrimination by making AI almost completely

193. Sullivan, *supra* note 15, at 428 ("In short, the current state of disparate impact law leaves the legality if [sic] Arti's operations unclear. At most, its use of explicit classifiers on prohibited grounds would be barred under a pure causal analysis, but its achieving much the same result by relying on factors correlated with but not formally race or sex may well be permitted."); *see also* Barocas & Selbst, *Big Data*, *supra* note 5, at 711–12; Bornstein, *supra* note 7, at 525.

194. *See supra* Section III.A.

useless relative to traditional statistical methods given the availability of so few variables.

This is the model of the ACA. The ACA inverts the traditional approach to combating rational discrimination from piecemeal removal of concerning traits to full scale removal of all traits, with limited exceptions. As a result, health insurers subject to the ACA are only allowed to consider four traits in their rating schemes: the number of people insured, their geographic area, their age, and whether they smoke.[195] The first two of these factors are not proxies for health because they are predictive of costs for reasons that are totally unrelated to health. By contrast, the latter two factors do indeed proxy for health, but in ways that reflect an intentional and considered policy judgment. By restricting insurance discrimination in rating to four pre-approved traits with well understood relations to the underlying suspect trait of health status, the ACA limits insurers' capacity to engage in proxy discrimination for policyholder health with AI by locating potential proxies, such as gym membership, eating habits, or medical debt. Indeed, the ACA model not only limits the potential for proxy discrimination for health,[196] but effectively limits proxy discrimination by gender, race, and other protected traits given the narrow scope of available traits to consider.

The ACA is not the only setting where this flipped default model has been employed. California's Proposition 103 is another example.[197] Under Proposition 103, auto insurers can only set premiums on the basis of an individual's driving record, mileage, years of driving experience, and "other factors that the commissioner may adopt by regulation."[198] Each of the law's pre-specified factors is predictive of risk for reasons that are orthogonal to legally suspect characteristics in auto insurance, like race and income. Similarly, Proposition 103's final discretionary category allows regulators to approve potential variables as long as those characteristics "have a substantial relationship to the risk of loss."[199] The system allows regulators to condition the inclusion of new factors into rates on insurers demonstrating that those

195. *See supra* note 87 and accompanying text. Some states restrict the allowable traits even further. *See* CTR. FOR CONSUMER INFO. & INS. OVERSIGHT, CTRS. FOR MEDICARE & MEDICAID SERVS., *Market Rating Reforms*, https://www.cms.gov/CCIIO/Programs-and-Initiatives/Health-Insurance-Market-Reforms/state-rating.html [https://perma.cc/E28F-N8SY].

196. *See, e.g.*, Jessica L. Roberts & Elizabeth Weeks Leonard, *What Is (and Isn't) Healthism?*, 50 GA. L. REV. 833, 845–46 (2016) (describing how the ACA limits health-status discrimination by insurers and "attempts to improve health insurance coverage").

197. *See* Dwight M. Jaffee & Thomas Russell, *Regulation of Automobile Insurance in California*, *in* AEI-BROOKINGS JOINT CTR. FOR REGULATORY STUDIES, DEREGULATING PROPERTY-LIABILITY INSURANCE: RESTORING COMPETITION AND INCREASING MARKET EFFICIENCY 195, 199 (J. David Cummins ed., 2002).

198. CAL. INS. CODE § 1861.02(a) (West 2003).

199. *Id.* Over time the California Insurance Commissioner has added a variety of optional rating factors, such as type of vehicle, completion of a driver training course, and, even, marital status of the driver. *See* CAL. CODE REGS. tit. 10, § 2632.5(d) (2019).

factors are predictive of risk for reasons having nothing to do with race, income, or other legally-suspect characteristics.[200]

Although flipping the default has been utilized in a few settings, it comes with significant efficiency and political economy tradeoffs. For instance, by limiting health insurers' capacity to leverage big data to help predict future claims experience, the ACA has caused some insurers' costs to outpace revenues, necessitating future premium increases for the entire pool.[201] At least partially as a result, the Trump Administration has adopted new policies that threaten to reopen the gates of proxy discrimination—or even intentional discrimination—by creating new exceptions to the ACA's strict limitations on health insurance discrimination.[202]

---

200. This mechanism has also been introduced in a narrower setting within genetic testing. In the United Kingdom, for example, an advisory committee was established to review which genetic tests insurers could take into account. *See* Prince, *supra* note 26, at 642–43.

201. *See, e.g.*, Tony Leys, *Iowa Teen's $1 Million-per-Month Illness No Longer a Secret*, DES MOINES REG. (May 31, 2017, 4:53 PM), https://www.desmoinesregister.com/story/news/health/2017/05/31/hemophilia-patient-costing-iowa-insurer-1-million-per-month/356179001 [https://perma.cc/BFF2-QZ8G]. This cycle of rising premiums is the so-called death spiral. It results when the increased premiums could result in individuals at lower risk opting to leave the insurance pool rather than take on costs disproportionately high for their associated risk. As more low-risk individuals leave the pool, the proportion of claims cost rises, resulting in another round of premium increases.

202. Certain health plans are reintroducing underwriting on the basis of multiple 'rational' characteristics back into the system. For example, in 2018, the Trump administration expanded the availability of short duration plans. Short Term, Limited Duration Insurance, 83 Fed. Reg. 38,212 (Aug. 3, 2018) (to be codified at 26 C.F.R. pt. 54, 29 C.F.R. pt. 2590, 45 C.F.R. pts. 144, 146, 148). There is ongoing litigation about the validity of these rules, but at the moment they remain valid. Katie Keith, *ACA Litigation Round-Up: Risk Corridors, CSRs, AHPs, Short-Term Plans, and More*, HEALTH AFF. (May 23, 2019), https://www.healthaffairs.org/do/10.1377/hblog20190523.823958/full [https://perma.cc/9PLR-8QTS]. Short duration plans were originally meant to be short-term stop gap insurance available to individuals as they transitioned between health plans, such as during a job transition. The plans are exempt from the ACA underwriting and coverage requirements and can therefore offer cheaper insurance to healthy individuals, although without offering coverage for many important healthcare needs. KAREN POLLITZ, MICHELLE LONG, ASHLEY SEMANSKEE & RABAH KAMAL, UNDERSTANDING SHORT-TERM LIMITED DURATION HEALTH INSURANCE 3 (2018), *available at* http://files.kff.org/attachment/Issue-Brief-Understanding-Short-Term-Limited-Duration-Health-Insurance [https://perma.cc/H93C-JGM3]. In early regulation, the short-duration plans were limited to less than three months. Excepted Benefits; Lifetime and Annual Limits; and Short-Term, Limited-Duration Insurance, 81 Fed. Reg. 75,316 (Oct. 31, 2016) (to be codified at 26 C.F.R. pt. 54, 29 C.F.R. pt. 2590, 45 C.F.R. pts. 144, 146, 147, 148); Sarah Lueck, *With Federal Rules Weakened, States Should Act to Protect Against Short-Term Health Plans*, CTR. ON BUDGET & POL'Y PRIORITIES: OFF THE CHARTS (Aug. 1, 2018, 11:00 AM), https://www.cbpp.org/blog/with-federal-rules-weakened-states-should-act-to-protect-against-short-term-health-plans [https://perma.cc/GH4D-3QTK]. The new Trump administration rules allow short duration plans to underwrite on the basis on pre-existing health conditions for policies that last up to 364 days, but that can be renewed for up to 36 months. Short-Term, Limited-Duration Insurance, 83 Fed. Reg. 38,212 (Aug. 3, 2018) (to be codified at 26 C.F.R. pt. 54, 29 C.F.R. pt. 2590, 45 C.F.R. pts. 144, 146, 148). While the short duration plans are allowed at the national level, some states are attempting to limit their scope. Lueck, *supra*; *see* Short-Term, Limited-Duration Insurance, 83 Fed. Reg. 38,212 (Aug. 3, 2018) (to be codified at

There has also been an industry shift to parsing risk (perhaps with the help of AIs) in marketing and product design, rather than rating. By so dramatically limiting insurer discrimination in rating, the law has arguably caused insurers to try to avoid high-risk customers in other ways, like developing a set of covered benefits that would be unattractive to those at higher risk or targeting marketing efforts to those at lower risk.[203] These trends highlight the invariable cat and mouse nature of addressing proxy discrimination, even with relatively aggressive legal tools.

Of course, the ACA has been a political lightening rod for a variety of reasons, not only due to its changes in rating. California's Proposition 103, for example, has been less controversial. However, given the radical changes that flipping the default brings, it is simultaneously one of the most effective strategies at combating algorithmic proxy discrimination and one that is perhaps the least likely to work.

It also has been implemented in two insurance contexts where the purchase of insurance is, or at least was intended to be, mandatory across a large risk pool. This therefore limits any impact of adverse selection. Should other areas, such as access to loans, housing, or employment, be similarly guaranteed no matter what one's traits—or upon consideration of only a specific few set of characteristics? The answer to this question is likely no in most settings. Perhaps a few other areas could be equally appropriate for a

---

26 C.F.R. pt. 54, 29 C.F.R. pt. 2590, 45 C.F.R. pts. 144, 146, 148); *see also* H.R. 1520, 29th Leg, Reg. Sess. § 431:10A (Haw. 2017); H.R. 2624, 100th Gen. Assemb., Reg. Sess. §§ 10, 15 (Ill. 2019); Maryland Health Care Access Act of 2018, H.R. 1782, 438th Gen. Assemb., Reg. Sess. §§ 6-102.1, 15-1202, 15-1301 (Md. 2018). In addition to short-duration plans, the Trump Administration has also increased the breadth and availability of association plans. Definition of "Employer" Under Section 3(5) of ERISA—Association Health Plans, 83 Fed. Reg. 28,912 (June 21, 2018) (to be codified at 29 C.F.R. pt. 2510). A recent district court opinion vacated these rules as "clearly an end-run around the ACA." New York v. U.S. Dep't of Labor, 363 F. Supp. 3d 109, 117 (D. D.C. 2019). However, this opinion has been appealed by the Trump Administration with an expedited review in the DC Circuit pending. Keith, *supra*.

203. *See, e.g.*, Marshall Allen, *Health Insurers are Vacuuming up Details About You—And It Could Raise Your Rates*, PROPUBLICA (July 17, 2018, 5:00 AM), https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates [https://perma.cc/6EGC-5JMC] (highlighting an insurer that used a square-dancing event as a way to attract healthy seniors to their pool, but also showing how cataloged data could help these marketing trends). The ACA anticipated these potential practices and prohibits insurers in the ACA marketplace from using marketing practices or benefit designs that "have the effect of discouraging the enrollment in such plan[s] by individuals with significant health need[]." 42 U.S.C.A. § 18031(c)(1)(A) (2010), *declared unconstitutional by* Texas v. United States, 945 F.3d 355 (5th Cir. 2019), *amended by* Pub. L. No. 116-94, 133 Stat. 2534 (2019) (Supreme Court cert. petition pending). But these restrictions do not apply to newer forms of health risk pooling. For example, there are no requirements that either association plans or short duration plans cover a minimum floor of essential health benefits, as there is with ACA health plans. Sarah Lueck, *3 Factors That Will Determine the Damage from Association Health* Plans, CTR. ON BUDGET & POL'Y PRIORITIES (July 27, 2018, 3:15 PM), https://www.cbpp.org/blog/3-factors-that-will-determine-the-damage-from-association-health-plans [https://perma.cc/4UG5-3Y9Z]; *see also* Essential Health Benefits Package ("EHB")—Benchmark Plan Standards, 45 C.F.R. § 156.110 (2015).

revolutionary change in access, such as higher education, but access to loans or other types of insurance are less likely to be viewed as social goods that justify such sweeping alterations of underwriting processes.

2. Expanding the Information Used: Requiring More Data to Limit Certain Types of Proxy Discrimination

Although prohibiting discrimination based on non-approved factors will naturally limit proxy discrimination by AIs, expanding the amount of information available to an AI could also decrease the occurrence of certain forms of proxy discrimination.[204] As recognized in the substantial literature on "statistical discrimination," much illegal discrimination is "rational" in the sense that it reflects real, statistical differences in relevant characteristics among different groups.[205] In many such cases, the discriminator could—with more effort—directly assess the relevant factor, rather than relying on illicit traits as a proxy for these factors. For instance, employers who have legitimate reasons to discriminate against individuals with criminal histories may "rationally" resort to discriminating on the basis of race given well-known race-based disparities in incarceration rates.[206]

As discussed in Part II, scenarios in which statistical discrimination is rational create the possibility that an AI may engage in proxy discrimination; a subtype that we labelled "indirect proxy discrimination."[207] But it also creates the very real possibility that AIs may, in fact, decrease the incidence of statistical discrimination—by proxy or otherwise—by reducing the costs of acquiring and processing data about directly relevant characteristics.[208] For instance, to the extent that past incarceration rates are indeed directly predictive of job performance for a particular employer, the AI might either

---

204. *Cf.* Strahilevitz, *supra* note 22, at 365–72 (suggesting that the government should publish more data about individuals, such as criminal history, in order to minimize racial discrimination).

205. *See generally* Charny & Gulati, *supra* note 42 (explaining that "statistical discrimination" occurs when the traits of a group serve as the basis of "employment decisions" instead of individual traits).

206. To appreciate this possibility, consider state efforts to pass "Ban the Box" legislation, which bars "employers from asking [applicants] about . . . criminal histories" in initial applications. *See* Agan & Starr, *supra* note 42, at 191. Although the goal of the legislation was to reduce barriers to employment for those with criminal convictions, the laws may have the unintended consequence of increasing racial disparity in hiring. *See id.* at 229. If an employer who does not want to hire anyone with a criminal record is prevented from collecting this information, they may instead turn to "statistical discrimination" strategies whereby they assume that black applicants are more likely to have a criminal record and therefore hire fewer blacks as compared to whites. Jennifer L. Doleac & Benjamin Hansen, *Does "Ban the Box" Help or Hurt Low-Skilled Workers? Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden* 4 (Nat'l Bureau of Econ. Research, Working Paper No. 22469, 2016).

207. *See supra* Section II.C.

208. *See supra* Section II.C.

be able to directly access this information or else to construct more reliable proxies than race for this information.

It follows that increasing AI's access to relevant data could decrease the program's need to rely on proxies for suspect characteristics by allowing it to more directly measure the factors that most directly relate to risk.[209] Adding data to AI models would also minimize situations where actors implicitly accept the possibility of decreased accuracy in their models to save costs on collecting or verifying information.[210]

At the same time, increasing the availability of data to AIs comes with many possible costs. First, while this strategy could decrease indirect proxy discrimination, it could also have the opposite effect. For example, providing more data to an AI that previously used race to proxy for a history of incarceration could cause it to derive a proxy for incarceration that omitted race, but it could also cause it to better target race so as to predict incarceration history. Second, and for similar reasons, increasing the AI's access to data would almost certainly increase the incidence of opaque and causal proxy discrimination, at least to the extent that legally suspect variables were directly predictive. Finally, increasing the availability of data comes at the expense of individual privacy. In the era of big data, employers, insurers, and lenders might have to access copious amounts of social, medical, and personal data, some mundane and some sensitive, about individuals to truly minimize indirect proxy discrimination. It is not clear that this is a trade-off worth making.

3. Transparency-Oriented Reforms

While the previous two solutions focused on the amount of data available to an algorithm, other possible solutions focus instead on how algorithms can or should employ that data. One such potential solution is to require discriminators to disclose information about how their algorithms impact members of protected groups.[211]

---

209. *See* Strahilevitz, *supra* note 22, at 368 (explaining that employers currently use proxies such as "spotty work history and being unemployed for more than a year" when the employer does not conduct a criminal background check). Although this will make it more difficult for some to access social goods. For example, although allowing employers to access information about criminal convictions will lower racial disparities in hiring, it will obviously not address legitimate public concerns about access to employment for those with a past conviction.

210. *See* Prince, *supra* note 26, at 651–52 (highlighting that insurers may be willing to trade some inefficiencies in modeling to save costs on data collection and verifying); *see also* Barocas & Selbst, *Big Data*, *supra* note 5, at 689–90 (discussing the costs of adding more data to algorithms and the resultant acceptance of less accurate models).

211. Hoffman, *supra* note 102, at 85. The European General Data Protection Regulation ("GDPR") also implemented requirements to provide information about and explain automated decision-making. Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841, 861 (2018).

Of course, this transparency-oriented approach would only allow third-parties to identify generalized disparate impact, rather than the more specific problem of proxy discrimination.[212] At the same time, robust disclosure regarding the impact of discrimination by AIs could help third-party researchers, litigants, and government entities to identify the subset of AIs that are most likely to be proxy discriminating. This is for two reasons. First, proxy discrimination necessarily produces a disparate impact, even if not all practices producing a disparate impact amount to proxy discrimination.[213] Disclosure of an AI's impact on protected groups can thus help third-parties isolate potential instances of proxy discrimination. Second, proxy discrimination will generally produce a distinctive type of disparate impact: The greater the statistical link between a legally protected characteristic and a facially neutral objective, the greater the magnitude of any disparate impact resulting from proxy discrimination.[214] This pattern should once again help provide red flags of proxy discrimination.

Consider an illustration of how those armed with appropriate data might be able to identify potential proxy discrimination. Suppose two similar large employers rely on AI and big data to guide their interviewing and hiring decisions. One of those employers offers robust employer-sponsored health insurance, while the other directs its employees to purchase coverage on the individual market, perhaps with the support of employer-funding through a Health Reimbursement Account.[215] Data showing that the first employer, which offered full health insurance to employees, also happened to hire substantially fewer individuals who had previously undergone genetic testing, would be highly suggestive that its AI was engaging in proxy discrimination. Not only would it show that the AI disparately impacted those likely to have a genetic condition, but it would do so in a context where there is likely to be a strong link between a legally protected characteristic (genetic information) and a facially neutral objective (reducing the costs of employer-sponsored health insurance).

Standing alone, this solution would merely increase the likelihood that firms using AIs to proxy discriminate could be publicly identified. Whether that result would help limit the prevalence of proxy discrimination would depend on a variety of factors, including the prospect that such information

212. *See id.* at 843–44 (proposing that actors provide counterfactuals to explain how their algorithm made a decision).

213. *See supra* Part II.

214. *See supra* Part II.

215. *See* Amy Monahan & Daniel Schwarcz, *Will Employers Undermine Health Care Reform by Dumping Sick Employees?*, 97 VA. L. REV. 125, 130 (2011). The Trump Administration recently released rules allowing employers to contribute pre-tax dollars to Health Reimbursement Accounts, which could then be used by employees to purchase coverage in the individual marketplace. *See generally* Health Reimbursement Arrangements and Other Account-Based Group Health Plans, 83 Fed. Reg. 54,420 (proposed Oct. 29, 2018) (to be codified at 45 C.F.R. pts. 144, 146, 147, 155).

could trigger negative media attention, new regulatory scrutiny, or novel legal theories. In all likelihood, however, this type of transparency reform would need to be paired with one of the more aggressive interventions described in subsequent Sections in order to meaningfully prevent proxy discrimination.

Moreover, there are a variety of different concerns and design issues that would come along with the collection and release of data regarding how protected groups fare when they interact with firms that deploy predictive analytics.[216] For instance, should the data be made available only to regulators, or also to the public? In either event, can the data be anonymized to reduce the likelihood that impacted individuals can be identified, particularly when their underlying membership in a protected group is potentially private information, as in the case of genetic information? Finally, what would be the challenges and costs of implementing this type of disclosure regime? Providing an explanation of algorithms can be complex and may run into other legal frameworks of trade secrets and privacy laws.[217]

These challenges are not, however, insurmountable, as demonstrated by the existence of exactly this type of disclosure regime in the home mortgage context. In particular, the Home Mortgage Disclosure Act ("HMDA") requires most lenders to report and make publicly available geocoded information regarding home loans, loan applications, interest rates, and the race, gender, and income of loan applicants.[218] This disclosure regime has promoted a massive amount of academic research and helped to identify both lending practices that disparately impact protected groups as well as intentional proxy discrimination in the form of redlining.[219]

4. Ethical Algorithms that Explicitly Control for Proxy Discrimination

While it is not possible to ex ante identify all potential proxies an AI may use,[220] it is possible to verify that specific characteristics are not proxies for suspect characteristics: Doing so simply requires showing that a characteristic

216. *See, e.g.*, Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robsinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 638–39 (2017) (highlighting various reasons why transparency-related reforms will not be successful).

217. Wachter, Mittelstadt & Russell, *supra* note 211, at 881–83.

218. Home Mortgage Disclosure Act of 1975, 12 U.S.C. § 2803 (2012); 12 C.F.R. §§ 203.4–203.5 (2018).

219. *See, e.g.*, DEBBIE GRUENSTEIN BOCIAN, WEI LI, CAROLINA REID & ROBERTO G. QUERCIA, CTR. FOR RESPONSIBLE LENDING, LOST GROUND, 2011: DISPARITIES IN MORTGAGE LENDING AND FORECLOSURES 31 (2011), *available at* https://communitycapital.unc.edu/files/2011/11/Lost-Ground-2011.pdf [https://perma.cc/PLS2-H9KT] (finding "that low-income and minority borrowers and neighborhoods have been disproportionately impacted by foreclosures and that this reflects the higher incidence of higher-risk products received by these groups"); Jacob S. Rugh & Douglas S. Massey, *Racial Segregation and the American Foreclosure Crisis*, 75 AM. SOC. REV. 629, 644–46 (2010) (finding a higher number and rate of foreclosures in metropolitan areas where there is a large "degree of Hispanic and especially black segregation").

220. For further discussion, see *supra* Section IV.B.1.

remains similarly predictive of outcomes even when controlling for membership in a suspect group. Proxy discrimination can therefore be eliminated from statistical models—whether or not those models are produced by AIs—through a conceptually straightforward statistical process.[221] The specifics of this process, as well as a range of more technical details, are described extensively in an important, though little appreciated, economics paper published in 2011.[222]

Counterintuitively, the first step in this process is for the statistical model under consideration to be re-estimated in a way that explicitly includes data on legally prohibited characteristics. For a model produced by an AI, accomplishing this requires including in the training data information on legally prohibited characteristics, such as the race or health status of individuals in the training population. This first step is necessary because it removes any predictive power that derives from legally permitted variables' capacity to proxy for a prohibited characteristic. In a model that explicitly includes all suspect variables, non-suspect variables will be treated as predictive only to the extent that they are predictive for reasons having nothing to do with their correlation to prohibited characteristics.

Having stripped from all permitted variables any predictive power attributable to proxy effects, the next step in the statistical process is to remove from the model any individualized information about legally prohibited characteristics. This step ensures that the ultimate model does not discriminate based on legally prohibited characteristic. Unfortunately, however, simply stripping the prohibited characteristic from the model can undermine the remainder of the model. Instead, therefore, it is generally necessary for the ultimate model to include consideration of the prohibited characteristic, but for it to assign the population average of that variable to every person subject to the model.[223]

---

221. *See* Pope & Sydnor, *supra* note 9, at 206–09. For practical proposals to implement ethical algorithms in insurance, see Birny Birnbaum, Presentation at CAS Ratemaking Seminar: Insurance Regulation: The Challenge of Big Data in Insurance (2018) (on file with Author).

222. According to Google Scholar, the Pope and Sydnor paper has been cited only 23 times since publication in 2011. Only two of those citations were from law reviews, and only one of them from a law review that was not co-authored by one of the co-authors of this paper. The paper's technique was recently modeled in the context of food safety and eating establishments. *See generally* Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias Into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 J. INSTITUTIONAL THEORETICAL ECON. 98 (2018) (discussing the limitations of the Pope & Sydnor applicability). Altenburger and Ho caution against using this technique in all settings; however, they specifically note that the technical solution may be inapt when attempting to address disparate impacts due to bias. This is different than the problems outlined in this paper where the protected trait is statistically linked to the desired outcome.

223. The process is actually more complicated for non OLS models, a matter which is addressed extensively in the Pope and Sydnor paper. The basic intuition, however, is the same across all types of statistical models: to ensure that "only the coefficients from the non-sensitive

Although this procedure for eliminating proxy discrimination in statistical models is reasonably straight-forward conceptually, it is also admittedly fraught with practical difficulties. In particular, not only does this approach require firms to collect data about legally prohibited characteristics, but it also requires them to attempt to measure the actual predictive power of these characteristics. It is easy to imagine how this process could unintentionally increase intentional discrimination; if a discriminator learned that a legally suspect characteristic was highly predictive, then it might be more inclined to intentionally discriminate on this basis. Moreover, this process may ironically have the effect of creating some of the very expressive harms that are generally absent from proxy discrimination; whereas proxy discrimination stealthily targets members of protected groups, the statistical process described above explicitly measures protected groups in a way that could conceivably produce some dignitary and communicative harms.[224]

Nor is it entirely clear that this statistical process would be legally permissible. In a very real sense, the process explicitly discriminates with respect to membership in a legally protected group in order to prevent the effects of such discrimination from being felt by these individuals. If this process were, for instance, legally mandated in an effort to prevent proxy discrimination, one could easily imagine a constitutional challenge suggesting that the government was forcing private actors to discriminate on the basis of sensitive characteristics.[225]

Even apart from these practical and legal difficulties, the costs associated with mandating the statistical maneuvers described above could potentially be substantial. Although the statistical approach is not complicated conceptually, it could well become immensely complicated as a practical matter, especially for the types of statistical models that AIs typically concoct. Nor is it clear that government regulators would have the technical expertise to ensure that this process was correctly performed and not manipulated for illicit ends.

Despite these very real concerns, using statistical methods to strip predictive models of the power to proxy for suspect characteristics represents one promising approach to combatting the emerging risk of proxy discrimination by AIs.

---

predictors are used when producing individuals' predicted values." Pope & Sydnor, *supra* note 9, at 207.

224. *See supra* Section III.B.4 (discussing how proxy discrimination does not likely produce the communicative harms of stereotyping, since most people do not know that they are in fact being stereotyped as a result of such discrimination).

225. Kroll et al., *supra* note 216, at 679–82; *cf.* Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 193 (2017) ("[A] simple prohibition on the use of protected characteristics such as race and sex in an automated decision process is easy to implement, but would do little to prevent biased outcomes. In any sufficiently rich dataset, proxy variables likely exist that closely correlate with these characteristics, permitting implicit sorting on those bases." (footnote omitted)).

5. Requirement of Potential Causal Connections

Algorithms and AIs rely on correlation, not causation—the predictive model identifies variables that are associated with the desired outcome without attempting to explain why such connections exist. Indeed, this is one of the very reasons that proxy discrimination is a likely phenomenon. The model does not care that the link between the variable and the desired outcome is due to association with a protected class; it only seeks to find the link. Indeed, because a model's goal is to find the best possible predictors though correlation, it will often be difficult, if not impossible, to determine from the model alone whether proxy discrimination is occurring.

One possible solution is to require those employing algorithms to convince regulators or others of a causal connections between the variables utilized and the desired outcome.[226] When a variable is causally linked to the desired outcome, it cannot be acting as a proxy for a protected trait. Consider, for example, the use of facial analysis by life insurers.[227] The AI could rely on many variables appearing in photographs. For example, the model could charge more for coverage to applicants whose photographs show stained teeth, indicating that they are likely to smoke. It is also, possible, however, to imagine an AI that utilizes features correlated with race, such as skin color or hairstyle, if underlying claims data shows a difference in mortality rates amongst whites and blacks. Requiring life insurers to establish a potential causal story would help to minimize proxy discrimination within the predictive facial modeling. It would stretch the imagination to derive a theory of causality between an applicant's hairstyle or skin color to mortality. In contrast, there is more plausible causality between baggy eyes and mortality —or at least one can describe a short series of causal links between lack of sleep and life expectancy that does not include race as part of the causal theory. The New York Department of Financial Services recently implemented such a causality requirement when life insurers discriminate on the basis of external data not collected from the policyholder.[228]

---

226. This approach to addressing the risk of proxy discrimination is also suggested by Grimmelmann and Westreich. Grimmelmann & Westreich, *supra* note 9, at 170 ("We believe that where a plaintiff has identified a disparate impact, the defendant's burden to show a business necessity requires it to show not just that its model's scores are not just correlated with job performance but explain it." (emphasis omitted)).

227. Barbara Marquand, *How Your Selfie Could Affect Your Life Insurance*, USA TODAY (Apr. 25, 2017, 10:03 AM), https://eu.usatoday.com/story/money/personalfinance /2017/04/25/how-your-selfie-could-affect-your-life-insurance/100716704 [https://perma.cc/MQL2-LZL4].

228. *See, e.g.*, Letter from James Regalbuto, Deputy Superintendent–Life Insurance, New York State Department of Financial Services, to All Insurers Authorized to Write Life Insurance in New York State (Jan. 18, 2019), *available at* https://dfs.ny.gov/industry_guidance/circular_letters/cl2019_01 [https://perma.cc/5GA9-KMSX] (warning that unfair discrimination laws can be implicated when "there is no demonstrable causal link between [a variable] and . . . mortality"). Like most states, New York both broadly prohibits unfairly discriminatory rates and specifically bars insurers from using protected traits such as race, national origin, past history

This solution is not without its challenges. Causality is not easy to identify,[229] and may be even more difficult to assess within the complex algorithmic environment.[230] For this reason, this solution to proxy discrimination by AI should not require definitive proof of causality, but rather a plausible causal story.[231] Additionally, it should not be expected that a comprehensive theory of causality be established. Some models use upwards of 70,000 variables to help predict desired outcomes. Establishing and assessing potential causality stories for each of the variables would be Herculean.[232] Instead, regulators might require plausible causal explanations for the subset of variables on which the AI most heavily relies.

Finally, some models will produce a score without indicating what variables have been utilized. With the predictive facial assessment, for example, machine learning may simply learn which faces are 'good' risks and 'bad' risks without indicating that hairstyle and teeth color are variables used in the calculations. Therefore, in some situations it may be necessary to implement a causality solution in conjunction with transparency requirements or ethical algorithm requirements. Once these tools identify the variables most highly-correlated with protected traits, assessment of causality can be narrowed to these variables. Despite the potential challenges, a causality requirement has the ability to limit proxy discrimination and increase perceptions of fairness in predictive models.[233]

---

of domestic violence. N.Y. INS. LAW § 2606 (McKinney 2015). Based on concerns of how growing use of algorithms and predictive models would challenge or circumvent these insurance anti-discrimination laws, the NY State Department of Financial Services launched an investigation into use of algorithms and external data in underwriting. The circular letter was an outcome of this investigation.

229. *See, e.g.*, Jill Gaulding, Note, *Race, Sex, and Genetic Discrimination in Insurance: What's Fair?*, 80 CORNELL L. REV. 1646, 1681 (1995) (noting that "[c]ausality . . . is a normative conclusion"); *see also* Austin, *supra* note 40, at 562 (arguing that the distinction between direct cause and indirect association "is inexact, if not entirely specious").

230. Indeed, in some cases the value of algorithms and machine learning is identifying relationships that are outside the bounds of human intuition. Selbst & Barocas, *Intuitive Appeal*, *supra* note 14, at 1094.

231. Gaulding, *supra* note 229, at 1681; Wortham, *supra* note 93, at 380 (arguing that variables can be fair when they "seem grounded in a causal explanation").

232. Wachter, Mittelstadt & Russell, *supra* note 211, at 853–54 (arguing that "the best tools for uncovering systematic biases are likely to be based upon large-scale statistical analysis and not upon explanations of individual decisions" and that establishing causal models will be both difficult and possibly "irrelevant").

233. Gaulding, *supra* note 229, at 1674, 1684. Gaulding establishes a "merged theory of fairness" that combines elements of anti-discrimination theories and what Gaulding calls, efficient discrimination—a concept that links to rational discrimination. *Id.* The merged theory holds that it is fair to use a variable correlated with a risk factor except if it is highly suspect and does not seem to be causally connected to the risk factor. *Id.* Additionally, using variables that cause, or likely cause, an outcome may be seen as more socially acceptable because these variables are more likely to be within the control of the individual, and causality has historically been "accepted [as a] basis for . . . assign[ing] . . . moral responsibility." Austin, *supra* note 40, at 559.

V. CONCLUSION

The emerging risks posed by AIs and big data have been the subject of innumerable law review articles, policy papers, articles in the popular press, books, and research articles in subject matters ranging from Philosophy to Computer Science to Sociology. Yet the precise ways in which AI and big data fundamentally change the risk of proxy discrimination are rarely laid out clearly in this vast literature, and quite frequently affirmatively misunderstood or totally ignored. This Article has demonstrated that AI and big data are game-changers when it comes to the risk of proxy discrimination, which—left unchecked—poses the prospect of undermining the core goals of all anti-discrimination regimes that seek to prohibit "rational" forms of statistical discrimination.

But the risk of proxy discrimination by AIs need not be left unchecked. To the contrary, policymakers have at their disposal a range of options for combatting these risks. While the most aggressive of these options would indeed substantially undermine the potential benefits of AI and big data, numerous less aggressive options are available that can allow for an appropriate balancing of the costs and benefits of emerging technologies. Rather than simply ignoring the accelerating threat of proxy discrimination by AIs, policymakers should confront this threat head-on in a way that reflects an informed and sober discussion of how to safeguard the advances made by existing anti-discrimination regimes.