BIAS PRESERVATION IN MACHINE LEARNING: THE LEGALITY OF FAIRNESS METRICS UNDER EU NON-DISCRIMINATION LAW

Sandra Wachter, Brent Mittelstadt,** and Chris Russell****

Abstract

Western societies are marked by diverse and extensive biases and inequality that are unavoidably embedded in the data used to train machine learning. Algorithms trained on biased data will, without intervention, produce biased outcomes and increase the inequality experienced by historically disadvantaged groups. Recognizing this problem, much work has emerged in recent years to test for bias in machine learning and AI systems using various fairness and bias metrics. Often these fairness metrics address technical bias, but not the underlying cause of inequality: social bias. In this Article we make three contributions. First, we assess the compatibility of fairness metrics used in machine learning against the aims and purpose of EU non-discrimination law. We show that the fundamental aim of the law is not only to prevent ongoing discrimination, but also to change society, policies, and practices to "level the playing field" and achieve substantive rather than merely formal equality. Based on this, we then propose a novel classification scheme for fairness metrics in machine learning based on how they handle pre-existing bias and thus align with the aims of non-discrimination law. Specifically, we distinguish between "bias preserving" and "bias transforming" fairness metrics. Our classification system is intended to bridge the gap between non-discrimination law and decisions around how to measure fairness in machine learning and AI in practice. Finally, we show that the legal need for justification in cases of indirect discrimination

---

* Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, UK. Email: sandra.wachter@oii.ox.ac.uk

** Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, UK.

*** Amazon Web Services, Inc. A great thank you is owed to the Harvard Law School, its faculty and students, the participants of the Harvard Law Faculty's Workshop, and the members of the Berkman Klein Center for Internet & Society for the inspiring discussions during Wachter's research visit in Spring 2020. The authors are also indebted to Dr Silvia Milano, Dr Johann Laux, and Prof Philipp Hacker for their detailed and immensely valuable feedback that greatly improved the quality of the paper. This paper would not exist without Jade Thompson, thank you for opening eyes, hearts, and minds, for caring and making others care. The paper also benefitted significantly from the insightful comments and diligent work of the editorial team at West Virginia Law Review. This work of the Governance of Emerging Technologies research programme at the Oxford Internet Institute has been supported by the British Academy Postdoctoral Fellowship grant nr PF2\180114 and grant nr PF\170151 from the Omidyar Group, and Miami Foundation.

can impose additional obligations on developers, deployers, and users that choose to use "bias preserving" fairness metrics when making decisions about individuals because they can give rise to prima facie discrimination. To achieve substantive equality in practice, and thus meet the aims of the law, we instead recommend using bias transforming metrics. To conclude, we provide concrete recommendations including a user-friendly checklist for choosing the most appropriate fairness metric for uses of machine learning and AI under EU non-discrimination law.

I. INTRODUCTION

Jade had always dreamt of studying mathematics at the University of Cambridge. In July 2020 her dreams were close to becoming reality. With her striking past record she was confident that she would reach the required marks on her final A-level exams. The COVID-19 pandemic had different plans. Due to the public health risks, in-person exams could not be held. Instead, an algorithm was used to predict the exam grade that she would have received based on her prior track record. Unfortunately, Jade's hopes were disappointed as the

algorithm predicted a low grade, apparently closing the doors to Cambridge.

In 2020 many high achieving students in England were punished by a standardisation algorithm designed to predict grades for A-level exams amidst the COVID-19 pandemic.[1] In an attempt to match historical distributions, the algorithm increased predicted grades at small, private schools and lowered grades at larger, state-run schools that have historically educated a larger proportion of Black, Asian and Minority Ethnic ("BAME") students.[2] As a result, BAME and poorer students disproportionately saw their predicted grades lowered compared to their peers.

Politicians and media were quick to point to a clear technical failure needing to be fixed. But it is worth asking the question: did the system actually fail, or did it perform precisely as designed?

The "Ofqual algorithm," like many algorithmic systems, was built on a very simple premise. Algorithms are designed to look at the past, find patterns, and predict the future.[3] Prior hiring decisions inform future hiring,[4] past loan and insurance decisions are the basis for future banking strategy and decisions,[5] past

---

1 The algorithm was designed by Ofqual, the country's regulator of qualifications, exams, and tests, to combat inflation in grades predicted by pupils' teachers. *See* Alex Hern, *Ofqual's A-level Algorithm: Why Did It Fail To Make the Grade?*, GUARDIAN (Aug. 21, 2020), https://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels.

2 Timandra Harkness, *How Ofqual Failed the Algorithm Test*, UNHERD (Aug. 18, 2020), https://unherd.com/2020/08/how-ofqual-failed-the-algorithm-test/.

3 Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI*, 2019 COLUM. BUS. L. REV. 494 (2019); Sandra Wachter, *Data Protection in the Age of Big Data*, 2 NATURE ELEC. 6, 6–7 (2019) [hereinafter Wachter, *Data Protection*].

4 For more information on issues of AI used in the workplace, see JEREMIAS PRASSL, HUMANS AS A SERVICE (2018) (ebook). *See also* Jeremias Adams-Prassl, *What If Your Boss Was an Algorithm? The Rise of Artificial Intelligence at Work*, 41 COMP. LAB. L. & POL'Y J. 123 (2019); Mark Burdon & Paul Harpur, *Re-Conceptualising Privacy and Discrimination in an Age of Talent Analytics*, 37 U. NEW S. WALES L.J. 679 (2014); Amit Datta, Anupam Datta, Jael Makagon, Dierdre K. Mulligan et al., *Discrimination in Online Advertising: A Multidisciplinary Inquiry*, 81 FAT 1–15 (2018); Jeremias Prassl & Martin Risak, *Uber, Taskrabbit, and Co.: Platforms as Employers? Rethinking the Legal Analysis of Crowdwork*, 37 COMPAR. LAB. L. & POL'Y J. 619 (2015).

5 For information on algorithms, bias, and credit, see generally Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019). *See also* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Tal Z. Zarsky, *Understanding Discrimination in the Scored Society*, 89 WASH. L. REV. 1375 (2014); Talia B. Gillis, *False Dreams of Algorithmic Fairness: The Case of Credit Pricing* (July 31, 2020), https://www.ssrn.com/abstract=3571266.

shopping history will impact future offers and prices,[6] previous tenants look like future tenants,[7] and the sentences of past criminals inform risk profiling for potential parolees.[8] Decisions about school admissions are no exception.[9] Like all algorithmic decision-making systems, the Ofqual algorithm was fed with historical data of not just Jade's past exam results, but also the results of past students at her school (and others), in order to predict her future.

The algorithm seems to have worked as designed. It adjusted the results of individual cases to match outcomes with historical data, namely the performance of schools in past A-level exams. This design also seems intuitively sensible at first glance and is the basis for many algorithmic systems in society: looking at how successful past students have been at university will seemingly give a reliable indication of how comparable students will perform in the future. It would appear we have "ground truth" from the past that can paint a reliable picture of the future.

If measured solely in terms of reproducing historical trends the Ofqual algorithm would appear well designed and reliable. Accuracy of this sort is often

---

[6] For information on AI bias when offering goods and services, see Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995 (2013). *See also* Karen Levy & Solon Barocas, *Designing Against Discrimination in Online Markets*, 32 BERKELEY TECH. L.J. 1183 (2017); Tal Zarsky, *Privacy and Manipulation in the Digital Age*, 20 THEORETICAL INQUIRES IN LAW 157 (2019). In view of price discrimination, see Oren Bar-Gill, *Algorithmic Price Discrimination: When Demand Is a Function of Both Preferences and (Mis) Perceptions*, HARV. JOHN M. OLIN DISCUSSION PAPER SERIES 18–32 (2018); Maurice E. Stucke & Ariel Ezrachi, *How Pricing Bots Could Form Cartels and Make Things More Expensive*, HARV. BUS. REV. (Oct. 27, 2016), https://hbr.org/2016/10/how-pricing-bots-could-form-cartels-and-make-things-more-expensive; Frederik Zuiderveen Borgesius, *Algorithmic Decision-Making, Price Discrimination, and European Non-Discrimination Law*, EUROPEAN BUS. L. REV. (forthcoming 2019).

[7] For information on bias in online housing markets, see Conference, *Auditing Race and Gender Discrimination in Online Housing Markets*, 14 ICWSM 24–35 (2020). For information on the 2008 housing crisis during which algorithmic tools were used, see JOSEPH E. STIGLITZ, THE PRICE OF INEQUALITY 239–42 (W.W. Norton ed., 2012).

[8] For more on algorithmic bias and policing, see Alexander Babuta & Marion Oswald, *Data Analytics and Algorithmic Bias in Policing*, ROYAL UNITED SERV. INST. (2019). *See also* Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 153–63 (2017); Rosamunde Van Brakel & Paul De Hert, *Policing, Surveillance and Law in a Pre-Crime Society: Understanding the Consequences of Technology Based Strategies*, 20 TECHNOLOGY-LED POLICING 163, 165 (2011). Regarding this issue in India, see Vidushi Marda & Shivangi Narayan, *Data in New Delhi's Predictive Policing System*, FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 317 (2020), https://doi.org/10.1145/3351095.3372865.

[9] For more discussion on bias and EdTech, see generally Elana Zeide, *The Structural Consequences of Big Data-Driven Education*, 5 BIG DATA 164 (2017). *See also* Priscilla M. Regan & Jolene Jesse, *Ethical Challenges of Edtech, Big Data and Personalized Learning: Twenty-First Century Student Sorting and Tracking*, 21 ETHICS INFO. TECH. 167 (2019); Elana Zeide, *The Limits of Education Purpose Limitations*, 71 U. MIAMI L. REV. 494 (2017).

the sole measure of performance in algorithmic systems. But is such an approach *fair*?

The Ofqual algorithm did not malfunction; rather, the design of the system resulted in technical bias that was not proactively identified and corrected. Specifically, predicted marks were based on the distribution of marks from a school over the previous year.[10] As a result, high performing students at high performing schools received high marks, whereas high performing students at low performing schools had their marks capped by the prior performance of other students and received a lower mark than deserved. In practice, this trend disproportionately affected BAME students.

Following public outcry this technical bias was quickly remedied. Marks were adjusted upwards according to teachers' predicted mark for individual students.[11] Fixing this technical bias did not, however, address the underlying social bias and inequality that contributed to certain schools underperforming historically.

In reality, the data used to study the past and predict the future was biased. It mirrored society as it exists, for better or worse. The cards were always stacked against Jade. The data reflected the effects of longstanding historical inequalities in access to good education, tutoring, giftedness programs, funds to supplement school provided resources, and parental support with schoolwork. Access to these educational necessities is heavily biased along racial[12] and gender[13] lines, the effects of which are reflected in past exams data.

Biases may have likewise been present but unacknowledged across Jade's educational journey. Standardised tests, including intelligence tests, have been shown to be racially biased against minority groups including Roma, Black

---

[10] David Hughes, *What is the A-level Algorithm? How the Ofqual's Grade Calculation Worked — and Its Effect on 2020 Results Explained*, iNEWS (Aug. 17, 2020), https://inews.co.uk/news/education/a-level-algorithm-what-ofqual-grades-how-work-results-2020-explained-581250.

[11] Harkness, *supra* note 2.

[12] The following examples have the sole purpose of showcasing some of the diverse inequalities faced by certain communities. The cited literature is taken from a wide variety of sources including cases from the United States and EU. The authors recognize that social barriers and inequalities manifest differently across different countries, and that specific barriers found in certain countries (e.g., the United States) cannot easily be presumed to occur in others (e.g. the UK). In fact, it is of great importance to assess inequality against the backdrop of the culture and history of a country to ensure that a locally accurate and comprehensive picture of existing inequality can be drawn. The examples offered here are solely intended to illustrate that seemingly objective and neutral data can reflect deep social inequalities and that heightened attention must be paid to the individual and collective social story behind the data points used to train machine learning and AI and make decisions in practice. JEAN HALLEY, AMY ESHLEMAN & RAMYA MAHADEVAN VIJAYA, SEEING WHITE 120–21, 127, 136 (2011) (ebook).

[13] ANGELA SAINI, INFERIOR 9–11 (2017).

people, Latinx, and migrants.[14] Encouragement and assessments of Jade's academic merit by her teachers and professors may also have been shaded by racial[15] and gender bias,[16] and ultimately reflected in her marks or reference letters from educators.[17] Similar gender biases may have influenced her male

---

[14] In regards to the Roma people and the EU, see D.H. v. Czech, 2007-IV Eur. Ct. H.R. 241, ¶¶ 1–122, https://hudoc.echr.coe.int/eng?i=001-83256. Standardised testing has a negative effect on children (i.e., being placed in special schools) and can significantly impact a particular minority if the class is composed of 50-90% Roma children. *Id.* This is seen as discriminatory due to Roma people only making up 2% of the general population. *Id.* at ¶ 134. *See also* Isabelle Chopin, Catharina Germaine & Judit Tanczo, Eur. Comm'n Directorate-Gen. for Just. & Consumers Unit JUST/D2, Eur. Network of Legal Experts in Gender Equal. & Non-Discrimination, *Roma and the Enforcement of Anti-Discrimination Law*, at 13–18 (2017), https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=605239. In regards to the United States, Black, Latinx, and immigrants, see HALLEY ET AL., *supra* note 12, at 40. *See also* CLAUDE S. FISCHER, MICHAEL HOUT, MARTÍN SÁNCHEZ JANKOWSKI, SAMUEL R. LUCAS ET AL., INEQUALITY BY DESIGN 172–73 (Princeton Univ. Press, 1996).

[15] *See* RENI EDDO-LODGE, WHY I'M NO LONGER TALKING TO WHITE PEOPLE ABOUT RACE 66–67 (2020); Simon Burgess & Ellen Greaves, *Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities*, 31 J. LAB. ECON. 535 (2013) (a study finding that black Caribbean and Africans are underassessed compared to their white peers and Indians, Chinese and mixed white and Asians studends are overassessed compared to white peers). In regards to mathematics skills versus received marks and encouragement based on ethnicity, see Rickie Sanders, *Gender Equity in the Classroom: An Arena for Correspondence*, 28 WOMEN'S STUD. Q. 182–93 (2000).

[16] *See* Victor Lavy & Edith Sand, *On the Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases* (Nat'l Bureau of Econ. Rsch., Working Paper No. 20909, 2015), https://www.nber.org/system/files/working_papers/w20909/w20909.pdf (finding teachers' biases to have a positive effect on boys grades and a negative effect on girls grades); Sue Wilson, *Teachers' Gender Bias in Maths Affects Girls Later*, CONVERSATION (Feb. 24, 2015), http://theconversation.com/teachers-gender-bias-in-maths-affects-girls-later-37844. As above concerning mathematics skill versus grades, the boys' grades improved in the following years. Girls however received the opposite treatment resulting in long term implications for their occupational choices and salaries. Their grades went down over time and they were less likely to take up mathematics and science later on in life.

[17] These stereotypes are also reflected in the biased way recommendation letters are written for male and female candidates. Male candidates are often described as having innate "genius," being "brilliant" or "trailblazers," whereas women are often described as "hard-working," "a team player" or "very knowledgeable." Maggie Kuo, *Recommendation Letters Reflect Gender Bias,* AAAS (Oct. 3, 2016), https://www.sciencemag.org/careers/2016/10/recommendation-letters-reflect-gender-bias. *See also* CAROLINE CRIADO PEREZ, INVISIBLE WOMEN: EXPOSING DATA BIAS IN A WORLD DESIGNED FOR MEN 102 (2019); EQUALITY IN HIGHER EDUCATION 80 (Equality Challenge Unit, 2013).

peers;[18] children have been shown, for example, to already develop clear ideas of gender roles by the age of six.[19]

Jade's experience with the Ofqual algorithm is not abnormal. Western societies are marked by diverse and extensive biases and inequality that are unavoidably embedded in the data used to train machine learning. Algorithms trained on biased data will, without intervention, produce biased outcomes[20] and increase the inequality experienced by historically disadvantaged groups.[21]

Recognizing this problem, much work has emerged in recent years to address bias in machine learning and AI systems.[22] Many scholars urge for greater accountability in their design and usage.[23] Machine learning systems take

---

18 Daniel Z. Grunspan, Sarah L. Eddy, Sara E. Brownell, Benjamin L. Wiggins et al., *Males Under-Estimate Academic Performance of Their Female Peers in Undergraduate Biology Classrooms*, 11 PLOS ONE, Feb. 10, 2016, at 1, 11–13.

19 In an interesting study from 2003 researchers showed that if children are shown pictures of people doing chores like sewing or cooking, they connect these activities already with gender roles. *See* Carol Lynn Martin & Diane Ruble, *Children's Search for Gender Cues: Cognitive Perspectives on Gender Development*, 13 CURRENT DIRECTIONS IN PSYCH. SCI. 67, 68 (2004). Children at the age of five were three times more likely to misremember seeing a girl cooking and sewing even though the picture showed a boy. *Id.*

20 *See generally* VIRGINIA EUBANKS, AUTOMATING INEQUALITY (2018); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2017).

21 EUBANKS, *supra* note 20; O'NEIL, *supra* note 20.

22 EUBANKS, *supra* note 20; O'NEIL, *supra* note 20; Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter et al., *The Ethics of Algorithms: Mapping the Debate*, 3 BIG DATA & SOC'Y, July–Dec. 2016, at 1, 1–15; Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y, Jan.–June 2016, at 1, 1–12; Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCI. 183, 183–86 (2017); Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J. LAW & TECH. 1, 1–64 (2017); Timnit Gebru, Jamie Morgenstem, Briana Vecchione, Jennifer Wortman Vaughan et al., *Datasheets for Datasets*, 1–24 (Cornell University Working Paper No. 1803.09010v7, 2020), https://arxiv.org/abs/1803.09010; Margaret Mitchell, Simone Wu, Andrew Zalidvar, Parker Barnes et al., *Model Cards for Model Reporting*, *in* 2019 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 220–29, https://arxiv.org/pdf/1810.03993.pdf; Tal Zarsky, *The Trouble with Algorithmic Decisions: An Analytic Road Map To Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 SCI., TECH. & HUM. VALUES 118 (2016); Conference, *The Problem with Bias: Allocative Versus Representational Harms in Machine Learning*, SIGCIS (2017).

23 *See generally, e.g.*, VIKTOR MAYER-SCHÖNBERGER & THOMAS RAMGE, REINVENTING CAPITALISM IN THE AGE OF BIG DATA (Basic Books, 2018); KAREN YEUNG & MARTIN LODGE, ALGORITHMIC REGULATION (2019) (ebook); B. Bodo, N. Helberger, K. Irion, F. Zuiderveen Borgesuis et al., *Tackling the Algorithmic Control Crisis—The Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents*, 19 YALE J.L. & TECH. 133 (2017); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54 (2019); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017); Omer Tene & Jules Polonetsky, *Taming the Golem: Challenges of Ethical Algorithmic Decision-Making*, 19 N.C. J.L. & TECH. 125 (2017);

in data and produce outputs, such as decisions or classifications, based on learned rules or parameters.[24] While biases in machine learning have many sources, there are two important categories of problematic bias that we refer to as (1) technical bias and (2) social bias.[25]

Problems in applying machine learning can induce additional biases that are not present in the data used to train the system or make decisions; we refer to these failures as "technical bias."[26] Technical biases reflect a failure of supervised learning algorithms to predict outcomes with the same accuracy across different protected groups leading to an increase in skewed, inaccurate, or unequal outcomes when compared to the training data.

But as Jade's experience shows, not all biases in machine learning can be traced back to technical sources or design choices.[27] The Ofqual algorithm had a simplistic design and did not malfunction. Its failures owed to ignorance of historical inequality in society and England's educational system. Ultimately it was the ignorance of social bias that led to technical bias in the design of the system.

Ari Ezra Waldman, *Power, Process, and Automated Decision-Making*, 88 FORDHAM L. REV. 613 (2019); Jatinder Singh, Ian Walden, Jon Crowcroft & Jean Bacon, *Responsibility & Machine Learning: Part of a Process* (Oct. 28, 2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2860048; Conference, *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, DATA AND DISCRIMINATION: CONVERTING CRITICAL CONCERNS INTO PRODUCTIVE INQUIRY (2014), http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf; Jatinder Singh, Jennifer Cobe & Chris Norval, *Decision Provenance: Harnessing Data Flow for Accountable Systems*, 7 IEEE ACCESS 6562 (2019); Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851 (2019); Andrew Tutt, *An FDA for Algorithms*, 84 ADMIN. L. REV. 83 (2017); Conference, *Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining*, 2125–26 (2016), https://doi.org/10.1145/2939672.2945386 (last visited Jan 9, 2021); Indrė Žliobaitė, *Measuring Discrimination in Algorithmic Decision Making*, 31 DATA MINING KNOWLEDGE DISCOVERY 1060 (2017); Laurens Naudts, *How Machine Learning Generates Unfair Inequalities and How Data Protection Instruments May Help in Mitigating Them*, *in* DATA PROTECTION AND PRIVACY: THE INTERNET OF BODIES 71 (Hart Publ'g 2019).

[24] *See* S. C. Olhede & P. J. Wolfe, *The Growing Ubiquity of Algorithms in Society: Implications, Impacts and Innovations*, 376 PHIL. TRANS. R. SOC'Y., A, June 2018, at 3–4; Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felton et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 642–47 (2017).

[25] This simplified view is inspired by Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM TRANS. ON INFO. SYS. 330, 333–35 (1996). *See also* Mireille Hildebrandt, *The Issue of Bias: The Framing Powers of ML*, MIT PRESS (forthcoming 2020) (manuscript at 2–9).

[26] Technical bias can be caused by factors such as an inappropriate choice of algorithm, inadequate features, insufficient sample size, insufficiently diverse data, and data drift. *See* Friedman & Nissenbaum, *supra* note 25, at 334.

[27] *Id.* at 333–35.

Comparatively speaking, "social biases" are very difficult to "fix."[28] They are a matter of politics, perspectives, and shifts in prejudices and preconceptions that can take decades to change. Biased outcomes should be expected when systems are trained on data that accurately reflects social reality, meaning it captures the biases and inequalities that characterise modern societies.[29] Unequal outcomes are not necessarily a result of inaccurate or incomplete data; rather, they can be an accurate reflection of the biased and unequal world in which machine learning is used.

Here we are dealing with a societal problem rather than a technical one. Adding more data will paint a more accurate and nuanced picture of the unequal world we live in for an algorithm to learn from or make decisions about, but it cannot resolve the root cause(s) of inequality; only individual, societal, or institutional change can. This is a feature of many technical fixes deployed in "fair machine learning": they are a temporary fix for the symptoms, but not causes, of inequality in society.[30]

Recognizing this limitation, we are left with three possible responses to algorithmic bias and resulting social inequality. First, we can do nothing and allow things to get worse. This does not require an active choice; failing to consider bias or fairness in designing, training, and using an automated decision-making process is often enough.[31] Non-intervention frequently amplifies and widens existing inequalities in our society that have been learned by a model

---

28 We define social bias as any systematic preference to make positive decisions for one group of people (or class of objects) relative to another. *See infra* Part III. This definition roughly follows the taxonomy proposed by Friedman and Nissenbaum. *See* Friedman & Nissenbaum, *supra* note 25, at 334. Following their taxonomy, "social bias" can be understood as a type of "preexisting bias" that "has its roots in social institutions, practices, and attitudes." *Id.* at 332–34. However, breaking with their taxonomy, we attribute biases arising from both individual and societal sources as "social bias."

29 *Id.* at 334.

30 *See, e.g.*, CATHERINE D'IGNAZIO & LAUREN F. KLEIN, DATA FEMINISM (MIT Press, 2020); Timnit Gebru, *Oxford Handbook of Ethics of AI: Race and Gender* (Markus D. Dubber eds., 2020); Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, 86 LANGUAGE 43 (2019); Cynthia L. Bennett & Os Keyes, *What is the Point of Fairness? Disability, AI and the Complexity of Justice*, AI FAIRNESS FOR PEOPLE WITH DISABILITIES (2019); Conference, *Roles for Computing in Social Change*, *in* CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 252 (2020); Conference, *Interventions Over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment*, *in* CONFERENCE ON FAIRNESS, ACCOUNTABILITY AND TRANSPARENCY 62 (2018); Workshop, *A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, *in* PROCEEDINGS OF THE 52ND HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES (2019); Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso et al., *Questioning the Assumptions Behind Fairness Solutions* (Nov. 27, 2018) (Position Paper, NeurIPS 2018 Workshop), https://arxiv.org/pdf/1811.11293.pdf; Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour et al., *Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions* (Apr. 28, 2020), https://arxiv.org/pdf/1811.07867.pdf.

31 *See generally* EUBANKS, *supra* note 20; O'NEIL, *supra* note 20.

through exposure to data reflecting existing biases and inequalities. Second, we can rectify technical biases and maintain the status quo to try to ensure our systems do not make things worse. Third, we can acknowledge the fact that the status quo is often not neutral and instead use AI and statistical analysis to shed light on existing inequalities. This can serve as a starting point for technical remedies and policy interventions that help fix historical biases and inequalities moving forward.

To date, much work in fair machine learning has focused on the second option: fixing technical bias,[32] maintaining the societal status quo, and, in general, trying to ensure machine learning does not make society more biased or unequal than is already the case.[33] This is a laudable goal and will remain tremendously important, especially given the fact that non-intervention alone suffices to widen inequalities. This route helps ensure that this will not happen.

This type of response, as well as what we term "bias preserving" fairness metrics that use the status quo as a baseline, seem to find legal recognition in Europe. They align closely with a fundamental normative concept in EU non-discrimination law: formal equality. Metrics that align with formal equality (or equality of treatment) aim to reproduce historic performance (as captured by the data) in the outputs of the target model with equivalent error rates for each group. In doing so, they aim not to make society more unequal than the status quo. Unfortunately, using these metrics run the risk of drawing away attention from the underlying causes of historical inequalities and thus can shift focus away from fixing them.

In contrast, the third response (which is related to what we coin as "bias transforming"[34] fairness metrics) aligns with a different fundamental normative concept in EU non-discrimination law: substantive equality. According to

---

[32] Mitigating technical bias is a particularly attractive challenge to data scientists and machine learning specialists because they are inherently tractable problems. In practice they can be fixed by improving performance against particular subgroups by improving the quality of training data in terms of volume, variety, accuracy, or representativeness, or by using machine learning techniques which better account for bias. *See generally* Anna Lauren Hoffmann, *Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse*, 22 INFO. COMMC'N & SOC'Y 900, 907 (2019); Abigail Z. Jacobs & Hanna Wallach, *Measurement and Fairness*, FACCT (forthcoming Mar. 2021), http://arxiv.org/abs/1912.05511.

[33] Of course, this is not to say that a technical fix cannot erode inequality (e.g., collecting gendered medical data), but more often than not technical fixes only scratch the surface. Similarly, if machine learning is viewed merely as a neutral tool that can be used for better or worse, picking a seemingly neutral baseline such as maintaining the status quo seems intuitively sensible.

[34] Here we use "bias transforming" rather than "debiasing" to reflect the idea that, as we have argued elsewhere, bias and the complementary notion of fairness are contextual, and forms of bias that are acceptable in one context, may not be acceptable in another. As such, there is no singular debiased state but instead an explicit transformative decision as to what form of bias would be acceptable in a particular context or application. *See* Sandra Wachter, Brent Mittelstadt & Chris Russell, Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI (Mar. 3, 2020) (unpublished manuscript), https://papers.ssrn.com/abstract=3547922.

substantive equality or "*de facto* equality,"[35] true equality can only be achieved by accounting for historical inequalities which actively ought to be eroded. The status quo is not treated as a neutral starting point from which to measure equality in opportunities and results; rather, protected groups start from different points which are not equal. Bias transforming fairness metrics reflect this observation and offer a starting point for possible interventions to address structural inequality in society.

This Article makes three contributions. First, we distinguish between two possible fundamental aims of non-discrimination law, formal and substantive equality, which impose different obligations for developers, deployers, and users of AI, machine learning, and automated decision-making. Second, we propose a classification scheme for fairness metrics in machine learning based on how they handle pre-existing bias (i.e., "bias preserving" and "bias transforming" fairness metrics) and how well they align with the aims of non-discrimination law. Finally, we recognize that the legal need for justification in cases of indirect discrimination may create new obligations for developers, deployers, and users. Recognizing this need for justification, we argue that metrics which require an explicit choice to be made about which biases a classifier should inherit should be preferred for purposes of fair decision-making under substantive equality. To conclude, we propose concrete recommendations and a checklist for choosing the most appropriate fairness metric under EU non-discrimination law.

To make these contributions, we draw on theories of EU non-discrimination law to show that bias preserving metrics can potentially be problematic when used as a benchmark for fairness in automated decision-making because they only pursue formal equality, not substantive equality. The fundamental aim of EU non-discrimination law is not only to prevent ongoing discrimination but also to change society, policies, and practices to "level the playing field" and achieve substantive rather than just formal equality.

We argue that developers, deployers, and users should, whenever possible, give preference to "bias transforming" fairness metrics (in particular to transformative versions of conditional independence) when a fairness metric is used to make substantive decisions about people in contexts where significant disparity has been previously observed. These metrics align best with the fundamental aim of EU non-discrimination law: substantive equality. The law expects private as well as public actors to play an active role in this endeavour, even if—as we will show—the precise duties for different types of actors remains a topic of open debate.

Of course, the use of bias transforming metrics does not automatically resolve all possible issues under non-discrimination law. Any disparity that may occur using bias transforming metrics remains open to dispute and requiring legal justification. However, by making the implicit choice of which bias a

---

35 MARIANNE GIJZEN, SELECTED ISSUES IN EQUAL TREATMENT LAW 23 (2006).

classifier should exhibit more explicit, bias transforming metrics draw attention to the underlying causes of social inequality. In doing so, they enable critical dialogue to distinguish socially acceptable disparities from those which cannot be justified and must be remedied.

Before proceeding further, a brief contextual note is necessary. Our intention is not to suggest that bias preserving metrics are without merit. In this regard we can draw a distinction between the usage of fairness metrics in machine learning for (1) diagnostic, testing, or research purposes, for example to identify biases inherited by a model or emergent technical biases, and (2) as a basis for making fair decisions in practice. Testing for and mitigating technical bias remains a vital area of research. Similarly, when making decisions about individuals in cases where an explicit normative decision has not yet been taken as to which biases the system should exhibit, technical bias mitigation helps ensure that systems are not making things worse. In scenarios where ground-truth labels can be exactly known,[36] no bias exists,[37] or where systems need to be designed to replicate social bias (e.g., as a diagnostic tool), technical bias mitigation is sufficient to ensure that no additional biases are induced through the use of machine learning. Finally, in jurisdictions that only pursue formal equality, biass preserving metrics may also be acceptable for decision-making.

Bias preserving metrics are likewise not illegal to use for automated decision-making in the EU. However, their usage in this regard introduces avoidable legal risks for developers, deployers, and users compared to bias transforming metrics. We argue that the use of bias preserving metrics for decision-making in contexts where known and unjustified inequality exists can give rise to prima facie discrimination. Under indirect discrimination doctrine, this means that disparity and the usage of such metrics need to be objectively justified under the "proportionality test." Whilst the use of such metrics can still be legal in Europe, we recommend that system operators pro-actively provide an objective justification for the use of bias preserving metrics.

---

[36] With regards to texting, when developing and deploying machine learning systems, it is vital to understand and account for the behaviour of these systems. In this context, fairness measures should not just be seen as a battery of contradictory constraints that cannot be simultaneously satisfied exactly but rather as a set of measures that can help illuminate unexpected behaviour, pinpoint forms of systematic errors, and aid in debugging. Bias preserving measures should remain an essential part of this process. With regards to a lack of ground-truth labels, for example, predicting the outcome of a preassigned biopsy test.

[37] For many scenarios, machine learning systems are used to predict a positive outcome following a positive intervention, for example: "If given a loan, would you repay?" To gather this data without sampling bias, a randomized control trial would be needed. *See also* Mitchell et al., *supra* note 22, at 21.

II. FORMAL AND SUBSTANTIVE EQUALITY IN NON-DISCRIMINATION LAW

EU non-discrimination law prohibits two types of discrimination: direct and indirect discrimination.[38] Direct discrimination means that a person is treated less favorably based on a protected attribute[39] (e.g., race and ethnicity,[40] gender,[41] religion and belief, age, disability, or sexual orientation[42]) that they possess in matters of a protected sector (e.g., the workplace, provision of goods and

---

38 *See generally* MARK BELL, ANTI-DISCRIMINATION LAW AND THE EUROPEAN UNION (2002). *See also* Erica Howard, *EU Anti-Discrimination Law: Has the CJEU Stopped Moving Forward*, 18 INT'L J. DISCRIM. LAW 60, 60–81 (2018); Justyna Maliszewska-Nienartowicz, *Direct and Indirect Discrimination in European Union Law–How to Draw a Dividing Line*, 3 INT'L J. SOC. SCI. 41, 41–55 (2014).

39 DIRECT DISCRIMINATION, EUROPEAN INSTITUTE FOR GENDER EQUALITY, https://eige.europa.eu/thesaurus/terms/108 (last visited Feb. 19, 20210).

40 *See, e.g.*, Council Directive 2000/43, 2000 O.J. (L 180) 22–26 (EC).

41 *See generally* Directive 2006/54 of the European Parliament and of the Council of 5 July 2006 on the Implementation of the Principle of Equal Opportunities and Equal Treatment of Men and Women in Matters of Employment and Occupation (recast), 2006 O.J. (L 204) 23, 23 (EC); Gender Equal Access to Goods and Services Directive 2004/113/EC European Implementation Assessment, EUR. PARL. DOC. PE 593.787 (2017).

42 *See generally* Council Directive 2000/78 of 27 November 2000 Establishing a General Framework for Equal Treatment in Employment and Occupation, 2000 O.J. (L 303) 16, 16–17 (EC). On how AI creates new groups unaccounted for under the law, see Linnet Taylor, *Safety in Numbers? Group Privacy and Big Data Analytics in the Developing World*, *in* GROUP PRIVACY: NEW CHALLENGES OF DATA TECHNOLOGIES 13, 13–36 (Linnet Taylor et al. eds., 2017); Alessandro Mantelero, *From Group Privacy to Collective Privacy: Towards a New Dimension of Privacy and Data Protection in the Big Data Era*, *in* GROUP PRIVACY: NEW CHALLENGES OF DATA TECHNOLOGIES 139, 139–58 (Linnet Taylor et al. eds., 2017); Brent Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, 30 PHIL. & TECH. 475 (2017); LEE A. BYGRAVE, DATA PROTECTION LAW: APPROACHING ITS RATIONALE, LOGIC AND LIMITS (2002); Tal Z. Zarsky, *An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics*, 14 ISJLP 11 (2017); Sandra Wachter, *Affinity Profiling and Discrimination by Association in Online Behavioural Advertising*, 35 BERKELEY TECH. L.J. (forthcoming 2020) [hereinafter Wachter, *Affinity Profiling*]. On how AI fairness methodologies fail to adequately account for the socially constructed nature of groups such as race, see Alex Hanna, Emily Denton, Andrew Smart, Jamie Smith-Loud, *Towards a Critical Race Methodology in Algorithmic Fairness*, *in* PROCEDEEINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 501 (2020), https://doi.org/10.1145/3351095.3372826.

services).[43] Different groups receive different levels of protection.[44] Direct discrimination is grounded in the Aristotelian postulate of treating "like cases alike" and treating "different cases differently" unless there is an objective reason not to do so. Equality achieved on these terms is also called "formal equality" or the "merit principle."[45]

Formal equality is not guaranteed to create equality of opportunity. To achieve equality of opportunity in practice, it is first necessary to acknowledge that widespread, structural inequality exists. It is not just single bad actors who openly discriminate that contribute to inequality; rather, it is the legacy and the functionality of institutions built on historical inequality that seamlessly maintain and exacerbate inequalities and inhibit substantive equality of opportunity in practice.[46]

In the words of President Lyndon B. Johnson at his 1965 Howard University Commencement Address:

> You do not take a person who, for years, has been hobbled by chains and liberate him, bring him up to the starting line of a race and then say, "you are free to compete with all the others,"

---

[43] For details on scope and limitations, see Wachter et al., *supra* note 34; Philipp Hacker, *Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law*, 55 COMMON MKT. L. REV. 1143, 1155–67 (2018); Frederik J. Zuiderveen Borgesius, *Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence*, 24 INT'L J. HUM. RTS. 1572 (2020); DAGMAR SCHIEK, LISA WADDINGTON & MARK BELL, CASES, MATERIALS AND TEXT ON NATIONAL, SUPRANATIONAL AND INTERNATIONAL NON-DISCRIMINATION LAW (2007). For a comparative view, see Raphaël Gellert, Katja de Vries, Paul de Hert, Serge Gutwirth, *A Comparative Analysis of Anti-Discrimination and Data Protection Legislations*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 61 (2013). On the scope of the European Convention of Human Rights, see F. Zuiderveen Borgesius, *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*, UNIV. AMSTERDAM DIGIT. ACAD. REPOSITORY 5 (2018), https://dare.uva.nl/search?identifier=7bdabff5-c1d9-484f-81f2-e469e03e2360; Janneke Gerards, *The Discrimination Grounds of Article 14 of the European Convention on Human Rights*, 13 HUM. RTS. L. REV. 99 (2013); Mark Bell, *The Implementation of European Anti-Discrimination Directives: Converging towards a Common Model?*, 79 POL. Q. 36 (2008).

[44] For a discussion on group hierarchy, see Erica Howard, *The Case for a Considered Hierarchy of Discrimination Grounds in EU Law*, 13 MAASTRICHT J. EUR. & COMPAR. L. 445 (2006).

[45] EVELYN ELLIS & PHILIPPA WATSON, EU ANTI-DISCRIMINATION LAW 5 (2012); Christopher McCrudden & Sacha Prechal, *The Concepts of Equality and Non-Discrimination in Europe: A Practical Approach*, 2 EUR. NETWORK OF LEGAL EXPERTS IN THE FIELD OF GENDER INEQ. 1, 1–50 (2009). For a critique of the circular nature of the merit principle and the need to fill it with substantive aims and values, see Peter Westen, *The Empty Idea of Equality*, 95 HARV. L. REV. 537 (1982).

[46] Catharine A. MacKinnon, *Toward a Renewed Equal Rights Amendment: Now More Than Ever*, 37 HARV. J.L. & GENDER 569, 572 (2014). The murder of Stephen Lawrence in UK is an example of institutional racism caused by omission rather than purposeful action. *See* ELLIS & WATSON, *supra* note 45, at 3.

> and still justly believe that you have been completely fair. Thus it is not enough just to open the gates of opportunity. All our citizens must have the ability to walk through those gates. This is the next and the more profound stage of the battle for civil rights. We seek not just freedom but opportunity. We seek not just legal equity but human ability, not just equality as a right and a theory but equality as a fact and equality as a result.[47]

Providing people with equal access to opportunities (i.e., formal equality) is not equivalent to providing access adjusted for historical disparities and their enduring effects on protected groups. The latter, referred to as "substantive equality" of opportunity (or "*de facto* equality"[48]), cannot be achieved simply by ignoring protected attributes (e.g., race, gender, disability) and treating everyone the same going forward. A more active attitude that accounts for social and historical realities is required. Inequality must be viewed not as something which needs to be proven on an individual basis, but rather as a background "fact of life" for certain groups that should be taken for granted unless disproven.[49]

A useful distinction can be drawn between procedural and substantive equal opportunity.[50] Formal equal opportunity focuses on procedural aspects of equal resource allocation. This includes removal of obstacles that affect certain groups (e.g., word-of-mouth recruitment). While better than formal equality (e.g., treating everybody the same), it still does not dismantle inequalities (e.g., unfair access to education).

In contrast, substantive equal opportunity focuses on positive measures that "level the playing field" to enhance fair competition (e.g., education or family friendly measures) in order to challenge established access criteria (e.g., job requirements) that reinforce existing patterns of disadvantage.[51] Here the

---

47 President Lyndon B. Johnson, Howard University Commencement Address: "To Fulfill These Rights" (June 4, 1965), https://teachingamericanhistory.org/library/document/commencement-address-at-howard-university-to-fulfill-these-rights/. For an overview of the history, aims and limitations of U.K. and U.S. non-discirimination law, see Christopher McCrudden, *Institutional Discrimination*, 2 OXFORD J. LEGAL STUD. 303 (1982).

48 GIJZEN, *supra* note 35, at 23.

49 STIGLITZ, *supra* note 7, at 199; Hoffmann, *supra* note 32, at 904; EDDO-LODGE, *supra* note 15, at 63.

50 Bernard Williams, *The Idea of Equality, in* PHILOSOPHY, POLITICS AND SOCIETY 125, 125–26 (P. Laslett & W.G. Runciman eds., 1962). In favor of Williams, see Sandra Fredman, *Substantive Equality Revisited*, 14 INT'L J. CONST. L. 712, 723–24, 735 (2016); SANDRA FREDMAN, DISCRIMINATION LAW (2011).

51 Williams, *supra* note 50, at 125–26; FREDMAN, *supra* note 50, at 723–24, 735.

focus is more on restructuring society rather than the individual, even though formal equal opportunity as an interim step has tremendous value.[52]

A. Indirect Discrimination and Substantive Equality

The concept of indirect discrimination was created to achieve substantive equality in practice.[53] Indirect discrimination "helps to dismantle underlying power structures . . . as well as to identify areas where further action is needed in order to achieve true equality, e.g. social engineering."[54] Indirect discrimination is intended to help redistribute resources from the advantaged to the disadvantaged and to promote diversity in society.[55] It enables non-discrimination law to play a more active role in creating substantive equality by tackling subtle social and historical inequalities.[56]

Indirect discrimination occurs when an "apparently neutral provision, criterion or practice"[57] that does not relate to a protected attribute is applied to a population equally but poses a particular disadvantage to a protected group. For example, a minimal height requirement in a job advertisement is not a case of direct discrimination because height is not a protected attribute. However, a height requirement is highly likely to create an indirect particular disadvantage for women who are, on average, shorter than men.[58] Elsewhere, the authors have argued that indirect discrimination is the most likely type of discrimination to arise from AI, machine learning, and automated decision-making because of the reliance of these systems on inferences and proxies of target variables and protected attributes.[59]

Once a claimant establishes prima facie indirect discrimination in court, the burden of proof shifts to the alleged offender.[60] The alleged offender then has

---

52 EDDO-LODGE, *supra* note 15, at 184.

53 Christa Tobler, European Commission, *Limits and Potential of the Concept of Indirect Discrimination*, at 5 (Sept. 2008), https://www.tandis.odihr.pl/bitstream/20.500.12389/20645/1/05963.pdf.

54 *Id.* at 24 (citing GIJZEN, *supra* note 35, at 82; SCHIEK, *supra* note 43, at 327).

55 GIJZEN, *supra* note 35, at 136.

56 Marc De Vos, *The European Court of Justice and the March Towards Substantive Equality in European Union Anti-discrimination Law*, 20 INT'L J. OF DISCRIMINATION & L. 62, 72 (2020); Sandra Fredman, *Equality: A New Generation?*, 30 INDUS. L.J. 145, 161 (2001); Tobler, *supra* note 53, at 24.

57 This language is stated in all EU Non-Discrimination Directives. *See also* Christopher McCrudden, *The New Architecture of EU Equality Law After CHEZ: Did the Court of Justice Reconceptualise Direct and Indirect Discrimination?*, 2016 EUR. EQUAL. L. REV. 1, 3 (2016).

58 This example is inspired by German case around height requirements for pilots. *See* Wachter et al., *supra* note 34, at 12.

59 Wachter et al., *supra* note 34; Wachter & Mittelstadt, *supra* note 3.

60 For more on this, see Julie Ringelheim, *The Burden of Proof in Antidiscrimination Proceedings. A Focus on Belgium, France and Ireland*, 2019 EUR. EQUAL. L. REV. 49, 49 (2019).

two options: (1) argue that indirect discrimination has not, in fact, occurred; or (2) acknowledge the disparity but offer an objective justification. Justified indirect discrimination occurs when the alleged offender pursued a legitimate aim and the mechanisms used pass the "proportionality test," meaning they are both legally necessary and proportionate. For example, physical requirements can be justified as essential when hiring firefighters on the basis of safety even if they impose a particular disadvantage.[61]

Indirect discrimination differs from direct discrimination by acknowledging that the social hurdles, struggles, and factual differences facing protected groups must be taken into consideration.[62] Indirect discrimination acknowledges the differences between groups and postulates that they ought to be treated differently. This is true even if rectifying existing inequality requires positive discrimination towards another group, for example more favorable changes in part-time work and employment law, which can be justified under the proportionality test.[63]

1. Positive Action and Substantive Equality

Protection under indirect discrimination and the aims of substantive equality in the form of equal opportunity are similar, but not equivalent to, positive action (referred to as "affirmative action" in the United States), and should not be confused.[64] Positive action, whilst also a form of substantive equality, focuses solely on equality of outcomes.[65] Adjusting the outcomes of a

---

On the practical limitations, see Lilla Farkas & Orlagh O'Farrell, *Reversing the Burden of Proof: Practical Dilemmas at the European and National Level*, EUR. NETWORK OF LEGAL EXPERTS IN THE NON-DISCRIMINATION FIELD 5 (2014).

61 For an extensive overview of EU case law on reasons to justify discrimination, see Wachter, *Affinity Profiling*, *supra* note 42, at 46–54.

62 De Vos, *supra* note 56, at 72.

63 *Id.* at 74; Sandra Fredman, *Addressing Disparate Impact: Indirect Discrimination and the Public Sector Equality Duty*, 43 INDUS. L.J. 349, 363 (2014).

64 ELLIS & WATSON, *supra* note 45, at 176–77.

65 It is worth noting that the concept of substantive equality and disparate impact doctrine is controversial in the United States. For a discussion and issues of bias mitigation in the U.S. context, see generally Catharine A. MacKinnon, *Substantive Equality: A Perspective*, 96 MINN. L. REV. 1 (2011); MARTHA MINOW, IN BROWN'S WAKE: LEGACIES OF AMERICA'S EDUCATIONAL LANDMARK 20 (2010); Fredman, *supra* note 50, at 713; Thomas Nachbar, *Algorithmic Fairness, Algorithmic Discrimination*, FLA. ST. U. L. REV. (forthcoming 2021); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 726 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CAL. L. REV. ONLINE 164 (2016); Zachary C. Lipton, Alexandra Chouldechova & Julian McAuley, *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 8125–35 (2018); Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291 (2020); Michael Feldman, Sorelle

procedure to be shared equitably across relevant protected groups is sufficient to achieve equality of results; for example, ensuring a 50/50 split of Catholics and Protestant police officers in Northern Ireland is an example of legal positive action.[66] The decision-making procedure itself need not change.

This is where substantive equality of opportunity differs: It seeks to create fair procedures using decision-making criteria that account for historical inequalities. The aim is not merely to give an advantage to certain members of a disadvantaged group by giving them a better outcome. Rather, substantive equality of opportunity seeks to create a level playing field for all participants by defining decision-making procedures and criteria with historical inequalities in mind (e.g., lowering the weight given to recommendation letters or Grade Point Average). Substantive equality is satisfied when everybody starts "the race" from the same point (e.g., via equal access to education or healthcare), not only when a specific number of people from a certain group have won the race.[67] Indirect discrimination "diagnoses discrimination,"[68] but does not necessarily achieve equality of outcomes. EU non-discrimination law, in pursuing substantive equality, thus aims to systematically erode inequalities over time. It fully supports measures for equal opportunity as well as for positive action, even with some restrictions on the latter.[69]

B. Substantive Equality Is the Aim of EU Non-Discrimination Law

Many non-discrimination scholars agree on the need to move away from a formalistic view of equality and adopt a proactive strategy that acknowledges the differences between groups and achieves substantive equality through structural change.[70] As Ellis and Watson argue,

---

Friedler, John Moeller, Carlos Scheidegger et al., *Certifying and Removing Disparate Impact*, *in* PROCEEDINGS OF THE 21ST ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 259, 259–68 (2015); CATHARINE A. MACKINNON, BUTTERFLY POLITICS (2017). For seminal work on U.S. non-discrimination law, see MARTHA MINOW, MAKING ALL THE DIFFERENCE: INCLUSION, EXCLUSION, AND AMERICAN LAW (1990). One could also argue that soft tools such as "outreach mechanisms" fall under the term of positive action. *See* Chantal Davies, *Exploring Positive Action as a Tool To Address Under-Representation in Apprenticeships* 77, 53 (2019), https://www.equalityhumanrights.com/sites/default/files/research-report-123-positive-action-apprenticeships.pdf.

66 ELLIS & WATSON, *supra* note 45, at 7.

67 Fredman, *supra* note 50, at 723, 729.

68 Fredman, *supra* note 56, at 161.

69 For case law on this issue, see Christopher McCrudden, *Gender-based Positive Action in Employment in Europe: A Comparative Analysis of Legal and Policy Approaches in the EU and EEA*, EUR. NETWORK OF LEGAL EXPERTS IN GENDER EQUAL. & NON-DISCRIMINATION 1, 220 (2019).

70 Fredman explains that "a four dimensional principle is proposed: to redress disadvantage; to address stigma, stereotyping, prejudice and violence; to enhance voice and participation; and to accommodate difference and achieve structural change." *See* Fredman, *supra* note 50, at 713.

> [i]f the moral basis on which the law forbids discrimination is that there is a fundamental human right to be treated in the same way as other human beings, the aim must logically be to produce substantive equality . . . [i]n particular, it involves taking an active attitude to dismantling the obstacles which stand in the way of equality.[71]

Similarly, Fredman argues that this active attitude cannot be limited in its focus on rectification of historical injustices, but rather should aim to achieve equal distribution of social goods for all people.[72] To achieve equality in these terms, differences in "capabilities" between protected groups must be accounted for because not everybody has the same abilities to achieve their goals;[73] rather, the ability to achieve is affected by "economic opportunities, political liberties, social powers, and the enabling conditions of good health, basic education, and the encouragement and cultivation of initiatives."[74]

This view is seemingly shared by the ECJ. In 2018 the Court opened the door to horizontal applicability of the non-discrimination principle in Article 21 of the Charter of Fundamental Rights of the European Union.[75] Article 21 of the Charter of Fundamental Rights of the European Union (non-discrimination) is now seen as a general and fundamental principle of the European Union.[76]

Jurisprudence of the ECJ likewise affirms that substantive equality is the intended aim of non-discrimination law, and that differences between groups must be acknowledged to achieve substantive equality in practice.[77] In cases

---

71 ELLIS & WATSON, *supra* note 45, at 4.

72 Fredman, *supra* note 56, at 156, also refers to human diginity as a principle of equality law to prevent levelling down to achieve parity (i.e., treating everyone equally bad).

73 AMARTYA SEN, DEVELOPMENT AS FREEDOM (2001); MARTHA C. NUSSBAUM, WOMEN AND HUMAN DEVELOPMENT: THE CAPABILITIES APPROACH (2000).

74 SEN, *supra* note 73, at 5.

75 For evolving case law, see Case C-144/04, Werner Mangold v. Rüdiger Helm, 2005 E.C.R. I-9981; Case C-555/07, Seda Kücükdeveci v. Swedex GmbH & Co. KG., 2010 E.C.R. I–21; Case 109/88, Handels- og Kontorfunktionærernes Forbund I Danmark v. Dansk Arbejdsgiverforening, acting on behalf of Danfoss, 1989 E.C.R. I-03199; Case C-414/16, Vera Egenberger v. Evangelisches Werk für Diakonie und Entwicklung eV, ECLI:EU:C:2018:257 (Apr. 17, 2018); Joined Cases C-569/16 & C-570/16, Stadt Wuppertal v. Maria Elisabeth Bauer and Volker Willmeroth v. Martina Broßonn, ECLI:EU:C:2018:871 (Nov. 6, 2018).

76 Article 21 states that "[a]ny discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited." *See* EUROPEAN UNION, *Charter of Fundamental Rights of the European Union*, C 364/1 (2000).

77 *See* De Vos, *supra* note 56, at 81. Among other cases, De Vos cites the following cases as evidence that the ECJ sees substantive equality as the aim of the law Case C-54/07, Centrum voor gelijkheid van kansen en voor racismebestrijding v Firma Feryn NV, 2008 E.C.R. I–397, http://curia.europa.eu/juris/liste.jsf?language=en&jur=C,T,F&num=C-54/07&td=ALL (last

which meet the strict requirements of prohibition of direct discrimination and formal equality (e.g., no one is treated differently on the basis of sex), the law acknowledges that systemic inequality can still occur, albeit in an indirect and more subtle manner. These disparities are often the legacies and the symptoms of illegal (institutional) discrimination.

C. *Positive Duties and Requirements for Substantive Equality*

While legal scholars broadly agree that the aim of EU non-discrimination law is substantive equality, they disagree about how best to achieve the necessary structural, institutional, and societal change in practice. According to Fredman,[78] Fraser and Honneth,[79] Collins,[80] and Barnard,[81] the goal

---

visited Mar 24, 2019); Case C-83/14, CHEZ Razpredelenie Bulgaria AD v Komisia za zashtita ot diskriminatsi, 2015 E.C.R. I–480, http://curia.europa.eu/juris/document/document.jsf?docid=165912&doclang=EN (last visited Mar 26, 2019); Case C-167/97, Regina v Secretary of State for Employment, ex parte Nicole Seymour-Smith and Laura Perez, 1999 E.C.R. I–60, http://curia.europa.eu/juris/showPdf.jsf?text=&docid=44408&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=6007788; Case C-144/04, Werner Mangold v Rüdiger Helm, 2005 E.C.R. I-9981, http://curia.europa.eu/juris/document/document.jsf?text=&docid=185565&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=7600685; Case C-303/06, S. Coleman v Attridge Law and Steve Law, 2008 E.C.R. I–415, http://curia.europa.eu/juris/document/document.jsf?text=&docid=67793&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=6050215 (last visited Mar 26, 2019); Case C-414/16, Vera Egenberger v Evangelisches Werk für Diakonie und Entwicklung eV, 2018 ECLI:EU:C:2018:257, http://curia.europa.eu/juris/document/document.jsf?text=&docid=201148&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=6616732; Case C-104/09, Pedro Manuel Roca Álvarez v Sesa Start España ETT SA, 2010 E.C.R. I-08661, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62009CJ0104; Case C-157/15, Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v G4S Secure Solutions NV, 2017 E.C.R. I–203, http://curia.europa.eu/juris/document/document.jsf?text=&docid=188852&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=6030648 (last visited Mar 26, 2019); Case 152-73, Giovanni Maria Sotgiu v Deutsche Bundespost, 1974 ECLI:EU:C:1974:131, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A61973CJ0152 (last visited Feb 17, 2021); Case C-177/88, Elisabeth Johanna Pacifica Dekker v Stichting Vormingscentrum voor Jong Volwassenen (VJV-Centrum) Plus, 1990 ECLI:EU:C:1990:383, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A61988CJ0177 (last visited Feb 17, 2021); Catherine Barnard & Bob Hepple, *Substantive Equality*, 59 CAMBRIDGE L.J. 562 (2000).

[78] Fredman, *supra* note 50, at 732.

[79] NANCY FRASER & AXEL HONNETH, REDISTRIBUTION OR RECOGNITION?: A POLITICAL-PHILOSOPHICAL EXCHANGE 36–37 (2003).

[80] Hugh Collins, *Discrimination, Equality and Social Inclusion*, 66 MOD. L. REV. 16, 24 (2003).

[81] Catherine Barnard, *The Future of Equality Law: Equality and Beyond*, *in* THE FUTURE OF LABOUR LAW: LIBER AMICORUM FOR BOB HEPPLE 213–28 (2004).

of non-discrimination law is not just eliminating social economic disadvantage, but also to foster social inclusion, participation in the community, and solidarity. Conversely, De Vos criticises the legislature and the Court of Justice for the lack of clear definition of those substantive equality goals and aims.[82]

The practical goals of substantive equality remain debated, including questions such as

- What is the end goal of non-discrimination law? To rectify historical harms and combat traditional power hierarchies?[83] To achieve equality of distribution of goods for all? To accommodate diversity?[84]
- What role (passive or active) is expected of the regulator, the legislature and the private and public sector?[85]
- Should there be a practice and pre-emptive duty of the public and the private sector to dismantle inequality?[86]
- Can this happen at the expense of dominant groups, potentially leading to positive discrimination?[87]
- When can disparity be legally justified?[88]
- How should the law address intersectional discrimination?[89]

The extent to which the law imposes and should impose positive obligations on public and private actors to achieve substantive equality is a particularly difficult question.[90] Positive obligations could require actors to, for

---

82 De Vos, *supra* note 56, at 83.

83 *See* MacKinnon, *supra* note 65; Fredman, *supra* note 50; Catharine A. MacKinnon, *Substantive Equality Revisited: A Reply to Sandra Fredman*, 14 INT'L J. CONST. L. 739, 746 (2016); Sandra Fredman, *Substantive Equality Revisited: A Rejoinder to Catharine MacKinnon*, 14 INT'L J.CONST. L. 747, 751 (2016); Catharine A MacKinnon, *Substantive Equality Revisited: A Rejoinder to Sandra Fredman*, 15 INT'L J. CONST. L. 1174, 1177 (2017) (discussing whether framing equality around traditional power hierarchies or around multifaceted ways of oppression is better to support substantive equality).

84 *See also* Fredman, *supra* note 56, at 164–65.

85 Urs Gasser & Carolyn Schmitt, *The Role of Professional Norms in the Governance of Artificial Intelligence: Some Observations and Outline of a Framework*, BERKAMN KLIEN CENTER (April 25, 2019), https://cyber.harvard.edu/story/2019-04/role-professional-norms-ai-governance; *Jonathan Zittrain and Jack Balkin Propose Information Fiduciaries to Protect Individual Privacy Rights*, TECH. ACADS. POL'Y (Sept. 28, 2018), http://www.techpolicy.com/Blog/September-2018/Jonathan-Zittrain-and-Jack-Balkin-Propose-Informat.aspx.

86 *See generally* Sandra Fredman, *Making Equality Effective: The Role of Proactive Measures* (Univ. Oxford Legal Rsch. Paper Series, Working Paper No. 53/2010, 2010).

87 *See* Fredman, *supra* note 63, at 363.

88 Wachter, *Affinity Profiling*, *supra* note 42, at 46–54.

89 *See generally* Kimberle Crenshaw, *Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color*, 43 STAN. L. REV. 1241 (1990); Devon W. Carbado, Kimberle Crenshaw, Vickie M. Mays, *Intersectionality: Mapping the Movements of a Theory*, 10 DU BOIS REV. 303 (2013); *see also* Wachter et al., *supra* note 34, at 20 (showing on the lack of legal protection under EU law).

90 *See generally* Fredman, *supra* note 86.

example, actively promote equal opportunity or even redistribute resources or jobs.[91] Negative obligations could restrict an actor's ability to base decisions on criteria that are known to disproportionately disadvantage certain protected groups.[92]

While case law of courts in the EU and UK is clearly moving toward substantive equality, specific active duties have not yet been formulated except in a few narrow cases.[93] Nonetheless, several possible grounds for broad positive duties have been identified.

In the UK, for example, the justification defense and the public sector equality duty may together create a duty for public bodies to set pre-emptive measures (i.e., without litigation) to prevent indirect discrimination if there are reasons to suspect that illegal disparity may occur.[94] In general, proactive measures of the private and the public sector might be more fruitful than just a complaint-based system. Reflecting this, some Member States have chosen this option.[95] However, a clear and general legal duty for preventative, positive duty

---

91 ELLIS & WATSON, *supra* note 45, at 7.

92 Such obligations raise further questions. For example, should banks be able to grant loans based on the financial situations of the clients? Should banks be required to use different or additional decision criteria that would level the playing field and give marginalized groups access to financial services? Similarly, under what circumstances can disparity be justified? For example, should a bank be able to justify indirect discrimination by establishing that it was necessary to consider income to achieve a legitimate interest? A legitimate interest in this case would be to only grant loans to applicants that are able to repay them. This practice can serve the interests of both the bank (i.e., to have loans repaid) and the loan applicant (i.e., to not be given an unpayable loan), but will exclude certain people from the market.

93 De Vos, *supra* note 56, at 72–73. Fredman, *supra* note 50, at 723–24, 735 (highlighting cases dealing with discrimination based on disability recognize a duty for "reasonable accommodation").

94 Fredman, *supra* note 63, at 363. Fredman thinks that in relation to public bodies in the UK in cases where indirect discrimination is likely to occur, a duty to take pre-emptive measures (including the duty to restructure) exists, but is unsure of whether a general pre-emptive duty for public bodies to actively dismantle inequality under the justification defense also exists. *Id.*

95 Fredman also criticises the ineffectiveness of complaint based systems, the lack of protection for claimants, the low conviction rates, and the high social and economic costs for (often very slow) litigation as well as the high hurdles in relation to the burden of proof (e.g., access to evidence, lack of comparator). Fredman, *supra* note 86, at 1–5. Complaints often do not deter perpetrators. *Id.* Proactive measures on the other hand would benefit everyone and not just the claimant and would contribute to the systematic erosion of inequality. *Id.* (discussing that some of the Member States have implemented mandatory (for the private and public sector as well as for trade unions) and optional rules (e.g., with incentives)).

is not yet established. Nonetheless, De Vos,[96] Tobler,[97] and Fredman[98] have argued that an implicit duty based on the existence of the prohibition of indirect discrimination should be inferred from EU law for both the public and private sectors, regardless of whether the institution in question was responsible for the inequality.[99] Nonetheless, it remains an open question as to what precisely these duties should entail.

Fitting with this argument for substantial change, Fredman believes that in cases where illegal direct or indirect discrimination occurred the outcome should not just be individual compensation but a requirement of restructuring.[100] Individual compensation cannot bring about structural change if not paired with a duty to restructure. Italy and Ireland, for example, have regulations that stipulate that if a claim is successful, the perpetrator is required to remove the discriminatory practice. Other legal remedies to bolster the complaints-based system and lessen the burden on individuals include strengthening the oversight powers of equality bodies and collective redress mechanisms.[101]

When dealing with scarce resources this restructuring can of course mean that traditional benefactors lose out. However, EU law[102] and case law[103] support measures of equal opportunity as well as positive action (with some restrictions)[104] to support societal and systemic restructuring. This holds true even if restructuring means that historically dominant groups receive less

---

96 *See* De Vos, *supra* note 60, at 71 ("A limited duty of preventive positive action is therefore implicit in the prohibition of indirect discrimination."); *see also* MARC DE VOS, BEYOND FORMAL EQUALITY: POSITIVE ACTION UNDER DIRECTIVES 2000/43/EC AND 2000/78/EC 81 (2007).

97 LIMITS AND POTENTIAL OF THE CONCEPT OF INDIRECT DISCRIMINATION, *supra* note 53, at 92.

98 Fredman, *supra* note 56, at 164, 167; Fredman, *supra* note 50, at 735 ("So far as EU law is concerned, the race directive does not specifically require the imposition of positive duties, although they are of course permitted. It instead stops at requiring a body to promote equal treatment and the obligation to promote social dialogue and encourage civil dialogue. At the same time, a degree of positive action is permitted.").

99 Fredman, *supra* note 56, at 164; Fredman, *supra* note 50, at 735.

100 Fredman, *supra* note 56, at 163.

101 On removing the discriminatory practice, see Fredman, *supra* note 86, at 5. On oversight powers, see European Network of Equality Bodies, *Meeting the New Challenges to Equality and Non-Discrimination from Increased Digitisation and the Use of Artificial Intelligence*, https://equineteurope.org/wp-content/uploads/2020/06/ai_summary_digital.pdf (last visited Feb. 17, 2021). On collective redress, see Sara Benedi Lahuerta, *Enforcing EU Equality Law Through Collective Redress: Lagging Behind?*, 55 COMMON MKT. L. REV. 783 (2018).

102 Fredman, *supra* note 63, at 363.

103 De Vos, *supra* note 56, at 74–75.

104 *See* McCrudden, *supra* note 69, at 220 (discussing case law on this issue).

favorable treatment and (potential) positive discrimination occurs.[105] Both can be justified under the EU's proportionality test.[106]

III. BIAS PRESERVATION IN FAIR MACHINE LEARNING

The existence and precise requirements imposed by positive substantive equality duties for the public and private sector remain an open question, but one that is critically important to the field of machine learning. Developers and users in both the public and private sector may have a duty to promote substantive equality in decision-making aided or driven by machine learning and derived technologies. Moreover, fairness and bias are growing areas of research in machine learning, with increasing attention being given to the intersection of technical metrics of fairness with the law. In both cases, designing technical capacities to meet current and future legal requirements concerning substantive equality is prudent. Matching technical capacities to measure bias and inequality with the aim and duties associated with non-discrimination law is thus of critical importance for developers and users of machine learning and AI.

To this end, in this section we propose a classification scheme for fairness metrics in machine learning based on the fundamental legal distinction between formal and substantive equality. We distinguish between metrics based on their treatment of historical social bias which affects their ability to support substantive equality in practice. We define social bias as any systematic preference to make positive decisions for one group of people (or class of objects) relative to another. In this formulation bias is a neutral concept, whereas the effects of the bias can be normatively significant.

Bias often carries a negative connotation because of its abstract formulation or effects in a given decision-making context. For example, if a loan officer has a bias to give loans at a greater rate to men compared to women, we can find the end state (i.e., men having greater access to loans than women) normatively problematic for a variety of reasons, following basic theoretical distinctions in moral and political philosophy.[107] The bias could be rejected for bringing about negative consequences, or violating some fundamental ethical

---

[105] *See* De Vos, *supra* note 56, at 81 ("However strict its case law may be, the fact remains that the Court does make room in principle for positive discrimination beyond mere positive action, as an exception to formal neutrality.").

[106] Fredman, *supra* note 63, at 363.

[107] We could argue that the bias violates an ethical principle of gender equality. We could likewise find the consequences of such a bias problematic in practice; over time, the preference would lead to men having greater access to financial services than women and thus greater ability to start businesses, purchase property, or otherwise engage in the market. From a legal perspective, we could find the loan officer's bias problematic because it factored gender, a legally protected attribute, into decisions regarding access to goods or services.

principle, or simply contravening legal provisions against direct discrimination.[108]

As is typically the case in fair machine learning research, throughout this Article we discuss bias with a negative connotation. Specifically, we view certain social biases in past decision-making as problematic because of the inequality they have created between protected groups of people in Western society.[109] From this observation we argue that preserving these biases in machine learning models can be problematic. If one were to reject the argument that existing inequality is in fact a problem, then one could likewise reject the argument that preserving that bias in fair machine learning is problematic.[110]

A. Fairness Metrics and Non-Discrimination Law

The concept of "indirect discrimination" and the "proportionality test" connect EU non-discrimination law with contemporary notions of algorithmic fairness.[111] In particular, some of the tests used by the European Court of Justice and Member State courts to measure indirect discrimination match the metric of demographic parity from algorithmic fairness.[112]

Formally, the definition of demographic parity asserts that each protected group, meaning a group based on a protected attribute such as race or gender, should, if it receives *k%* of the positive decisions, then also receive *k%* percent of the negative decisions.

Building from this observation, Wachter, Mittelstadt, and Russell[113] and Wachter[114] examine the justification of apparent indirect discrimination, by which a practice that would otherwise be considered discriminatory can be justified on the basis of a legitimate interest and the proportionality test.[115]

---

108 *Supra* Part II.

109 More precisely, in this Article we treat social inequality as problematic because European non-discrimination law aims at substantive equality between groups. Inequality can, of course, also be criticised on many other legal, ethical, and political grounds.

110 Throughout this Article we assume that the reader finds at least some of the inequality that currently exists in the world normatively problematic.

111 *See* Wachter et al., *supra* note 34.

112 *Id.* at 48–54. We also emphasize that the courts do not expect decision-making systems to perfectly satisfy demographic parity. Rather, other points to be considered include the impact, or potential harm, of the decision on each individual; the number of people impacted by the system; and the size of the systematic violation of demographic parity. *Id.*

113 *Id.* at 41–44.

114 Wachter, *Affinity Profiling*, *supra* note 42, at 46–54.

115 *See id.* at 41–44 (detailing the ECJ case law on what can be objectively justified); *see also supra* SectionII.A. Further, examples can be seen in the application of the UK Equality Act 2010 that transposes the EU Non-Discrimination Directives. *See Words and Terms Used in the Equality Act*, EQUALITY & HUM. RTS. COMM'N https://www.equalityhumanrights.com/en/advice-and-guidance/commonly-used-terms-equal-rights (last visited Feb. 28, 2020). Here the UK Equality

Justifications can, in principle, be offered to defend systems that apparently cause indirect discrimination by violating demographic parity.

We have argued elsewhere that if justification is accepted as a defense this should imply that the system as a whole should satisfy the algorithmic fairness notion of Conditional Demographic Parity (also referred to as Conditional Independence in the statistical literature). This metric can be tested in practice. Elsewhere we have proposed a simple metric for robustly estimating the effect size of conditional systematic bias: Conditional Demographic Disparity ("CDD").[116]

A decision-making system is said to exhibit Conditional Independence, with respect to a particular protected attribute, such as race or sex, and a conditioning attribute, such as salary or length of employment, if

(1) any difference in how the system collectively treats people with a particular race or sex can be attributed entirely to differences in the conditioning attribute; and

(2) after conditioning on this variable, the decisions made are statistically independent of the protected group.

Much as conditional independence explicitly encodes a dependence on a deliberately selected conditioning attribute, a class of metrics we refer to as "bias preserving" should be recognized as an explicit dependence on target labels. As such, bias preserving metrics implicitly advance an answer to a difficult normative question: what is the right factor to depend on in a given use case? By answering this question, these metrics also assume that a single correct answer can be given for a wide range of challenging use cases. This characteristic of bias preserving metrics is particularly pronounced for equalized odds, which is formally defined as a special case of conditional independence that conditions on the target labels.[117] However, it also holds for other bias preserving metrics that, by matching error rates across groups, are in one form or another seeking to preserve the distribution of target labels.

---

and Human Rights Commission offers the following guidance on requirements for an objective justification that would satisfy the courts. *Id.* The justification must show that "the aim must be a real, objective consideration, and not in itself discriminatory (for example, ensuring the health and safety of others would be a legitimate aim); if the aim is simply to reduce costs because it is cheaper to discriminate, this will not be legitimate; working out whether the means is "proportionate" is a balancing exercise: does the importance of the aim outweigh any discriminatory effects of the unfavorable treatment?; there must be no alternative measures available that would meet the aim without too much difficulty and would avoid such a discriminatory effect: if proportionate alternative steps could have been taken, there is unlikely to be a good reason for the policy or age-based rule." *Id.*

116 Wachter et al., *supra* note 34.

117 *See* Moritz Hardt, Eric Price & Nati Srebro, *Equality of Opportunity in Supervised Learning*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 3315–23 (Daniel D. Lee et al. eds., 2016).

To this end, the application of indirect discrimination provides a model of key questions that should be asked and answered before algorithmic fairness metrics are used in practice. The questions are asked with an eye toward the future: we imagine a hypothetical scenario where a given algorithmic decision-making application is contested in court for causing indirect discrimination. In this context two key questions must be answered:

(1) Does significant disparity exist?

(2) Accepting that significant disparity exists, is it justified?

Developers and users of machine learning should ideally proactively answer these questions at the point of deployment with an eye towards future liability, but also to demonstrate a commitment to substantive rather than merely formal equality.

The first question concerns how we define relevant groups (i.e., disadvantaged and comparator groups), and how we measure disparity between them. For our purposes here we ignore the former which we have addressed in detail elsewhere.[118] With regards to the latter, the first question can be rephrased as: which fairness metric should we use to measure disparity? And, more specifically, which variable(s) should the test be conditioned on? This is a key normative question, as the answer can result in potentially problematic inequality being ignored or obscured from view, particularly when it is a result of past biases and inequalities.

B. Bias Preserving and Bias Transforming Fairness Metrics

To help answer these questions in the context of a legal framework designed for substantive equality, we define two types of fairness metrics. "Bias preserving" fairness metrics seek to reproduce historic performance in the outputs of the target model with equivalent error rates for each group as reflected in the training data (or status quo). In contrast, "bias transforming" metrics do not blindly accept social bias as a given or neutral starting point that should be preserved, but instead require people to make an explicit decision as to which biases the system should exhibit.

To formalise our notion of bias preserving fairness, we say that any fairness metric is *bias preserving* if it is always satisfied by a perfect classifier that exactly predicts its target labels with zero error, replicating bias present in the data. Fairness metrics that are not necessarily satisfied by a perfect classifier, we refer to as *bias transforming*.[119]

To understand how existing fairness metrics should be classified, we refine our hypothetical scenario to one where a machine learning system is trained to make decisions $\hat{Y}$, such as if an individual will be hired, based on

---

118 Wachter et al., *supra* note 34, at 13–32.

119 As a negative category, metrics classified as bias transforming will be less homogeneous than those classified as bias preserving.

historic target data Y. Y could come from various sources, for example, the system could be trained on historic data such as who was previously hired, or who passed probation and became a permanent employee. Metrics that can be classified as bias preserving, such as "equalized odds,"[120] equal opportunity,[121] and calibration,[122] implicitly assume that various forms of bias in the historic data Y are there for a reason and should be preserved.[123]

This treatment of the status quo and existing bias as neutral, or as something to be preserved, can be troubling for a variety of reasons. Past hiring decisions reflect the biases of past hiring managers, while the seemingly more objective criteria of passing probation introduces a collection of potential causes of bias. For example, failure to pass probation may reflect a hostile working environment, and of course such data can only exist for people that were previously hired by the hiring manager.

Returning to the fairness metrics in question, equalized odds is formally defined as: "predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y if $\hat{Y}$ and A are independent, conditional on Y," where $\hat{Y}$ is the output of a system, and Y, some input labels the system is trying to predict.[124]

Based on this definition, equalized odds is a form of conditional independence or conditional demographic parity, conditioned on historic data Y, and reflecting its biases exactly. The problem is that blind application of equalized odds locks in these historical biases without providing a justification for relying on the metric, and thus the historic biases, going forward. From the perspective of EU non-discrimination law, which requires justification when prima facie discrimination is established,[125] using bias preserving metrics in contexts where discrimination or significant, unjustified disparity has been previously established can cause a problem for developers, deployers, and users. Specifically, we suggest that doing so effectively provides potential claimants with evidence of prima facie discrimination and shifts the burden of proof to the alleged offender to justify the disparity.[126]

The relationship between conditional independence and other fairness metrics is not as mathematically exact. However, all these metrics have in common the idea that bias present in the target labels data is meant to be there,

---

120 Hardt, Price, & Srebro, *supra* note 117.

121 *Id.*

122 Chouldechova, *supra* note 8.

123 A full taxonomy of fairness metrics along with references is provided in Appendix 1, Table 1a.

124 Hardt, Price, & Srebro, *supra* note 117. Here we use equalized odds as an example because it is self-evidently bias preserving. Specifically, it is functionally a type of conditional independence that conditions on the target labels and thus preserves the labels' bias.

125 Wachter et al., *supra* note 34, at 41–44.

126 *Infra* Part V.

and a perfect classifier that exactly reproduces the given labels (i.e., $Y = \hat{Y}$) would satisfy all such metrics. As such, all of these metrics should be understood as trying to prevent machine learning systems from inserting new bias into a system by preserving the bias present in the data.[127] We refer to such fairness metrics as bias preserving.

This observation naturally raises a question: how common is bias preservation in fairness metrics proposed in the fair machine learning literature? A 2018 "state of the art" review identified 20 distinct metrics, 13 of which are bias preserving by definition.[128]

According to our definition, a fairness metric can be classified as bias preserving if a perfect classifier $Y = \hat{Y}$ is guaranteed to exactly satisfy the metric. In such cases, a decision about whether the biases present in the labelled data Y are acceptable should be made before the metric is used. This does not mean that the other fairness metrics would be appropriate to use, simply that other questions need to be asked.

Fairness metric	Bias preserving?
1. Group fairness, Statistical (demographic) parity	✗
2. Conditional statistical (demographic) parity, Conditional independence	✗
3. Predictive parity, outcome test	✓

[127] In behavioral economics terms, such metrics display "status quo bias," meaning their design reflects a preference for maintaining the status quo. *See* William Samuelson & Richard Zeckhauser, *Status Quo Bias in Decision Making*, 1 J. RISK & UNCERTAINTY 7 (1988).

[128] Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, 2018 IEEE/ACM INT'L WORKSHOP ON SOFTWARE FAIRNESS 1; *see also* Table 1. For further reading on the topic of different (competing) fairness definitions, see also Reuben Binns, *On the Apparent Conflict Between Individual and Group Fairness*, *in* 2020 PROCEDURES OF CONFERENCE ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 514; Sorelle A. Friedler, Carlos Scheidegger & Suresh Venkatasubramanian, On the (Im)possibility of Fairness (Sept. 23, 2016), https://arxiv.org/abs/1609.07236.pdf; Meike Zehlike, Philipp Hacker & Emil Wiedemann, Matching Code and Law: Achieving Algorithmic Fairness with Optimal Tansport (Sept. 24, 2019), https://arxiv.org/abs/1712.07924.pdf; Alice Xiang & Inioluwa Deborah Raji, On the Legal Compatibility of Fairness Definitions (Nov. 25, 2019) (Workshop paper, NeurIPS 2019), https://arxiv.org/abs/1912.00761.pdf; Martim Brandão, Helena Webb, Marina Jirotka & Paul Luff, *Fair Navigation Planning: A Resource for Characterizing and Designing Fairness in Mobile Robots*, 282 A.I. 103259 (2020); Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary et al., *A Comparative Study of Fairness-Enhancing Interventions in Machine Learning*, *in* 2019 PROCEDURES OF CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 329; Deborah Hellman, *Measuring Algorithmic Fairess*, 106 VA. L. REV. 56 (2020).

4. False positive error rate balance	✓
5. False negative error rate balance, Equal opportunity	✓
6. Equalized odds	✓
7. Conditional use accuracy equality	✓
8. Overall accuracy equality	✓
9. Treatment equality	✓
10. Test-fairness or calibration	✓
11. Well-calibration	✓
12. Balance for positive class	✓
13. Balance for negative class	✓
14. Causal discrimination (direct discrimination)	*
15. Fairness through unawareness	*
16. Fairness through awareness	✗
17. Counterfactual fairness	✗
18. No unresolved discrimination	✗
19. No proxy discrimination	✗
20. Path based causal reasoning	✗

Table 1 – Bias preserving fairness metrics

* Indicates that a perfect classifier satisfying $Y = \hat{Y}$ would always satisfy this definition if perfect predictions can be made without explicitly using the protected attribute such as race or sex. **N.B.** Formulas and references for each metric can be found in Appendix 1, Table 1a.

The fact that one classifier can satisfy multiple fairness metrics might be somewhat surprising given the variety of impossibility theorems that state that they are incompatible. However, these theorems explicitly exclude perfect classifiers as special cases.[129] As such, the differences and incompatibilities between different bias preserving fairness metrics should be understood as engineering decisions that alter how the system balances misclassification errors, but that does not change the suitability of a perfect classifier. In other words, the choice of bias preserving metrics is essentially a decision about how properties of the target variable distribution should be preserved by a new classifier. As a result, a perfect classifier that preserves all properties of the target distribution with zero error satisfies all of these metrics.

---

[129] Chouldechova, *supra* note 8; Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores (Nov. 17, 2016), https://arxiv.org/abs/1609.05807.pdf; Geoff Pleiss, Manish Raghavan, Felix Wu, John Kleinberg et al., On Fairness and Calibration (Nov. 3, 2017) (Conference paper, NIPS 2017), https://arxiv.org/abs/1709.02012.pdf; Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, Cass R. Sunstein et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113 (2018).

C. *Limits of Bias Preserving and Transforming Metrics*

Within the current literature, the key difference between bias transforming and bias preserving metrics is that most *bias transforming* metrics are satisfied by matching *decision rates* between groups, while *bias preserving* metrics typically require matching *error rates* between groups. For example, the bias transforming metric demographic parity is satisfied if positive decisions are made at the same rate across the relevant groups, e.g., if $x\%$ of both Black and white people receive positive decisions. Conditional demographic parity (conditioning on salary) is satisfied if $x\%$ of both Black and white people earning over a threshold receive positive decisions and also some $y\%$ of both Black and white people earning under the threshold receive positive decisions.

Other bias transforming fairness metrics include causal methods such as counterfactual fairness.[130] These methods explicitly model forms of societal bias using structured causal models in order to eliminate them. More sophisticated variants allow for explicit normative decisions as to which forms of bias should be preserved in the form of path-specific effects defined over the causal graph.[131]

In contrast, bias preserving metrics typically match error rate across groups. For example, equalized odds requires the ratio of true positive to false negative decisions to be the same across groups, and for the ratio of true negatives to false positives to also be matched across groups. In contrast, the technical metric "equal opportunity" (not to be confused with the legal notion of equality of opportunity discussed above) only requires the first of these constraints: the ratio of true positive to false negative decisions needs to be the same across groups.

This definition in terms of error rate inextricably links bias preserving metrics to the use of "ground-truth" labels. For example, while we can say if a deployed algorithm exhibits (conditional) demographic parity, without generating new ground-truth labels for the data coming in we cannot say whether it satisfies equalized odds, only that it did so on previous training or validation data. This dependence on ground-truth data for evaluation makes it particularly difficult to say whether a system satisfies a bias preserving fairness metric if distribution shift occurs between training the system and its deployment.

Generating ground-truth data at deployment for people that have received negative decisions is practically challenging. It is highly unlikely, for example, that banks would regularly give loans to applicants that fail their risk

---

130 Matt J. Kusner, Joshua Loftus, Chris Russell & Ricardo Silva, Counterfactual Fairness (Mar. 8, 2018) (Conference paper, NIPS 2017), https://arxiv.org/abs/1703.06856.pdf; Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt et al., Avoiding Discrimination through Causal Reasoning (Jan. 21, 2018) (Conference paper, NIPS 2017), https://arxiv.org/abs/1706.02744.pdf; Silvia Chiappa & Thomas P. S. Gillam, Path-Specific Counterfactual Fairness (Feb. 22, 2018), https://arxiv.org/abs/1802.08139.pdf.

131 Chiappa & Gillam, *supra* note 130.

assessment solely for the sake of creating the data necessary to know whether their decisions satisfy equalized odds, or are formally fair in practice. Data does not exist for the counterfactual needed to assess whether a rejected job or a loan applicant would have been successful in the event of a positive decision.[132] As a result, social "randomised control trials," where positive outcomes are assigned to random applicants, would often be the only way to determine whether a deployed system truly satisfies bias preserving metrics.

In the field of machine learning it is common in practice to train a machine learning system on proxy variables that are easier to measure than the variables we want the system to predict. For example, a system may be trained to predict if an individual has a high credit score as a proxy for if they will repay a loan, or a system may be trained to predict if an individual will be arrested as a proxy for if they will break the law. This mismatch between what we want to predict and the proxy variables we can actually observe is another way for systematic bias to enter systems.[133] Inheriting social bias in this way cannot be detected by naïve use of bias preserving metrics that simply measure whether the machine learning system recovers values of the proxy variables with similar errors for each groups.

Bias transforming metrics for algorithmic fairness can of course also be problematic. Such metrics require an explicit decision as to what biases a system should exhibit, and these decisions can interact with existing social biases in non-obvious ways. For example, blindly enforcing *demographic parity* when deciding who should receive a loan could result in loans being offered to individuals that are unable to repay the loans due to existing social biases resulting in lower salaries. In turn this could lead to more individuals in disadvantaged groups going bankrupt, and worsening social inequality. Further, the chosen conditions can also be highly political and normatively laden. Developers may purposefully choose favorable variables to condition on. Nonetheless these choices would—if published alongside summary statistics (which we have advocate for elsewhere)—be transparent and open to inspection and debate by affected parties and regulators.[134]

---

132 Mitchell et al., *supra* note 30, at 4.

133 Proxies can reflect different historical biases and institutional inequality than typically associated with the intended prediction variables. For example, assessing a student's potential based on standardised testing scores carries latent bias caused by disparity in access to education and support services (e.g., external tutoring). With regards to Roma and the EU, see D.H. v. Czech Republic, 2007-IV Eur. Ct. H.R. 241, https://hudoc.echr.coe.int/eng?i=001-83256. Standardised testing has a negative effect on children (i.e., being placed in special schools) and can significantly impact a particular minority if the class is composed of 50-90% Roma children. This is seen as discriminatory due to Roma people only making up 2% of the general population. For more details, see CHOPIN ET AL., *supra* note 14, at 13–18; FISCHER ET AL., *supra* note 14, at 172–73 With regards to the United States and Black and Latin people and immigrants, see HALLEY, ESHLEMAN, & VIJAYA, *supra* note 12, at 40, 120–21, 127, 136.

134 Wachter et al., *supra* note 34, at 62–64; *see supra* Section III.B.

Returning to non-discrimination law, the need to justify measures that cause significant disparity draws attention to the importance of choosing the right fairness metric in a given decision-making context. Choice of metric of course matters less for diagnostic, debugging, and investigatory purposes. However, when used as a basis to actually make fair (automated) decisions in practice, choice of fairness metric is of critical normative importance. Choosing an appropriate metric should be subject to significantly more explicit consideration and justification than is currently the case in work on fair machine learning.

Our proposed classification of metrics according to bias preservation is intended to help evaluate possible metrics and choose between them. Justification as required by EU non-discrimination law provides an ideal model to evaluate the acceptability of fairness metrics at a sectoral and local deployment level. In the following sections we explore the relative merit and possible justification of bias preserving and transforming metrics in relation to substantive equality.

IV. THE STATUS QUO IS NOT NEUTRAL

Algorithmic decision-making can only be neutral in a normative sense if we are satisfied with how decisions have been made in the past. Specific actions are not required to inherit and reinforce the biases of past decision-making. If we were solely transposing good or equitable human decision-making processes to automated systems, we would only need to worry about technical bias. This is, unfortunately, typically not the case.

Recognizing this, bias preserving fairness metrics are potentially problematic on several grounds. Many of their limitations can be traced back to their treatment of the status quo as a neutral starting point to assess fairness in machine learning. These metrics do not differentiate between reasons for past inequality; rather, only the replication of historic performance with comparable error rates for each group matters. Simply matching these error rates is considered "fair."As a result, they ignore underlying causes of social biases and inequalities in a given decision-making context.[135] In contrast, bias transforming metrics require a positive normative choice to be made about which biases should be exhibited by a decision-making system. In forcing this choice, any recognised instance of disparity between groups may be seen as potentially discriminatory and in need of legal justification.[136]

By design, bias preserving metrics run the risk of "freezing" or locking in social injustices and discriminatory effects, which does not align well with the core aim of EU non-discrimination law: to achieve substantive equality. Ignoring

---

135 HALLEY ET AL., *supra* note 12, at 120–21, 127, 136. Racist hiring practices, for example, may be indistinguishable from inequality rooted in broader societal factors, such as people of color having fewer educational opportunities and thus being less competitive in the job market.

136 *Supra* Section III.B.

the reasons behind inequality is problematic from the view of substantive equality because understanding why decisions were made historically is crucial to correct the inequalities they created.

In Western society, the status quo is not acceptable for large parts of the population. The way we make decisions is often marked by prejudice and inequality.[137] Historical trends in decision-making have led to diminished and unequal access to opportunities and outcomes among certain groups.[138] It is in this sense that the status quo is not neutral. Maintaining it by treating it as a neutral baseline for comparison cannot therefore be considered a politically, ethically, or legally neutral act.

Individual and institutional prejudice are ingrained in many countries.[139] Racial inequality in top jobs is particularly pronounced in the United States and

---

[137] Evidence cited in this section provides a small glimpse of the vast structural racism, sexism, ableism, heterosexism and other bigotries embedded in the data we collect and use to train decision-making algorithms. The brief overview provided here cannot possibly be comprehensive, and cannot do justice to all disparities in the world. Many other scholars have done the ground-breaking work necessary to understand discrimination, prejudice, and inequality as found in the 21st century. Rather, what follows is merely a sample of this inequality, focusing in particular on statistics and stories that reveal the significant disparity to be found in historical data and the status quo. Given the regulatory frameworks discussed in this work, and reflecting the frequent comparisons in literature on AI policy and regulation, we focus on evidence from the United States, UK, and EU. All categories and labels used for different demographic categories reflect those used in the primary sources cited.

[138] *See, e.g.*, ANGELA Y. DAVIS, WOMEN, RACE, AND CLASS (1983).

[139] In the UK, for example, significant proportions of the UK population have readily admitted to holding racist views, including a significant group comprised of highly educated and affluent white professionals between the ages of 35 and 44. *See* EDDO-LODGE, *supra* note 15, at 65, 190, 205; *30 Years of British Social Attitudes Self-Reported Racial Prejudice Data*, NATCEN SOC. RSCH., https://www.bsa.natcen.ac.uk/media/38110/selfreported-racial-prejudice-datafinal.pdf (last visited Feb. 28, 2021); Matthew Taylor & Hugh Muir, *Racism on the Rise in Britain*, THE GUARDIAN (May 27, 2014), https://www.theguardian.com/uk-news/2014/may/27/-sp-racism-on-rise-in-britain. Signs of institutional racism can also be found in self-reported attitudes; statistics from 2014, for example, revealed that "[c]oncern about immigrants as a drain on public service resources rises significantly with income, while job-related concern declines as income rises." *See* Duffy Bobby & Tom Frere-Smith, *Perceptions and Reality: Public Attitudes to Immigration*, IPSOS MORI SOC. RSCH. INST., at 56 (2014).

United Kingdom.[140] People of color can be treated worse in interviews,[141] can be less associated with higher paid occupations, can appear less qualified than their white counterparts with the same qualifications,[142] and can receive lower wages on average.[143]

Women and gender non-binary people face similar challenges in the job market.[144] Female associated jobs pay less on average and are seen as less

---

[140] Statistics from 2019 from the United States reveal 88.8% of chief executives are white, while only 4.1% are Black, 5.8% are Asian, and 6.2% are Hispanic. *See Employed Persons by Detailed Occupation, Sex, Race, and Hispanic or Latino Ethnicity*, U.S. BUREAU OF LABOR STATISTICS, https://www.bls.gov/cps/cpsaat11.htm (last visited Feb. 28, 2021). Similar worrying trends are seen in areas of general and operational management, advertising and promotion managers, sales and marketing managers, financial managers, and industrial production managers, where the vast majority (between 80 and 90%) of positions are filled by White people. In the UK, the situation is slightly better. A survey from 2018 shows roughly 11% of Indian, Asian, and White people are managers, directors, or senior officials. However, only 5% of Black people occupy these high-level roles. *See Employment by Occupation*, GOV'T UK (Nov. 9, 2020), https://www.ethnicity-facts-figures.service.gov.uk/work-pay-and-benefits/employment/employment-by-occupation/latest. "Elementary" jobs—the lowest skilled category of occupation recorded in the survey—was highest in the Black (16%) and Other White (15%) ethnic groups.

[141] *See* Carl O. Word, Mark P. Zanna & Joel Cooper, *The Nonverbal Mediation of Self-fulfilling Prophecies in Interracial Interaction*, 10 J. EXPERIMENTAL SOC. PSYCH. 109 (1974) (demonstrating how interviewers treat Black people differently than White people. For example, interviewers sat further away from Black candidates, spoke with more errors, were more unfriendly and ended the interview more quickly than for White candidates. When the same interviewing patterns were applied to White people, the performance declined and the White candidates were more nervous.); *see also* HALLEY ET AL., *supra* note 12, at 155–56.

[142] HALLEY ET AL., *supra* note 12, at 153.

[143] In-group differences also exist. In the United States, lighter-skinned African-American men receive roughly the same wages as White men, whereas medium and dark-skinned African-American men do not. Arthur H. Goldsmith, Darrick Hamilton & William Darity, *From Dark to Light: Skin Color and Wages Among African-Americans*, 42 J. HUM. RES. 701 (2007).

[144] For a discussion on gender bias, AI, and the workplace, see Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes et al., *Bias in Bios: A Case Study of Semantic Representation Bias in a High-stakes Setting*, *in* 2019 PROCEEDINGS OF CONFRENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 120 (2020). Society and institutions including schools and universities, politics, and the workplace have historically been built on a social expectation of heteronormativity. *See* JEAN HALLEY & AMY ESHLEMAN, SEEING STRAIGHT: AN INTRODUCTION TO GENDER AND SEXUAL PRIVILEGE 32–33 (2016); András Tilcsik, *Pride and Prejudice: Employment Discrimination Against Openly Gay Men in the United States*, 117 AM. J. SOCIO. 586 (2011). People identifying as LGBTQ, or not conforming with binary standards of heteronormativity, are afforded little legal protection across the world and routinely face severe inequality, harassment, discrimination, and violence. *See* Alex Hanna, *Being Transgender on the Job Market*, INSIDE HIGHER ED. (July 15, 2016), https://www.insidehighered.com/advice/2016/07/15/challenge-being-transgender-academic-job-market-essay. A 2014 U.S. report showed that one out of two transgender individuals are sexually assaulted or abused at some point in their lives. *See Responding to Transgender Victims of Sexual Assault*, OFF. OF JUSTICE PROGRAMS (June 2014), https://ovc.ojp.gov/sites/g/files/xyckuh226/files/pubs/forge/sexual_numbers.html.

important than male associated jobs.[145] Decisions about promotions and hiring are likewise prone to gender bias,[146] with women routinely rated as less competent than male colleagues despite comparable performance.[147] In the United States in 2019 only 27.6% of chief executives were women.[148] Physical appearance can also influence employer's assessment of the talent and professionalism of employees and applicants.[149]

Social stigma around disability also manifests in workplace inequality.[150] People with physical, psychosocial, intellectual, and sensory conditions routinely face significant challenges, biases, and inequalities in the workplace,[151] including a lack of employment, promotion, mentorship, and significantly higher likelihood of dismissal.[152] Mental conditions in particular are viewed more severely than physical conditions.[153] Ableist assumptions result in

---

[145] A study from the United States showed that as more women take up particular jobs the sector's perceived prestigiousness and average wages decline. *See* PEREZ, *supra* note 17. Women also face greater difficulty when negotiating remuneration due to not being seen as "likeable" in negotiations. *See* Hannah R. Bowles, Linda Babcock, Lei Lai, *Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes It Does Hurt To Ask*, 103 ORGANIZATIONAL BEHAVIOR & HUM. DECISION PROCESSES 84 (2007). Service jobs such as teaching, counselling, nursing, or childcare, which are often associated with women, tend to pay less than other occupations with similar requirements. *See* Paula England, Michelle Budig & Nancy Folbre, *Wages of Virtue: The Relative Pay of Care Work*, 49 SOC. PROBS. 455 (2002); *see also* HALLEY ET AL., *supra* note 12, at 160–61.

[146] O'NEIL, *supra* note 20, at 106–22. Criteria that seem prima facie neutral, such as consecutive years of employment, are in reality sexist as they punish career disruptions commonly experienced by women (e.g., caring duties).

[147] For example, in a U.S. study faculty assessors (both male and female) ranked female competence lower for a laboratory manager position and offered lower starting salaries even though identical resumes were sent in but some had male and some had female names. *See* Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brecoll, Mark J. Graham et al., *Science Faculty's Subtle Gender Biases Favor Male Students*, 109 PROC. NAT'L ACAD. SCI. 16474 (2012); *see also* SAINI, *supra* note 13, at 5.

[148] *Employed Persons by Detailed Occupation, Sex, Race, and Hispanic or Latino Ethnicity*, *supra* note 140.

[149] For example, a study shows that women who wear their hair in braids at work (which is common for women from India and African-Americans) are more likely to be seen as unprofessional in some workplaces. *See* HALLEY ET AL., *supra* note 12, at 150–51.

[150] For example, Harpur argues that people with disability are often seen as ugly, in need of "curing" (including eugenics), and are encountered with charity and pity. PAUL DAVID HARPUR, ABLEISM AT WORK 7 (2019). For an overview of the cause of this stigma, see MICHELLE R. NARIO-REDMOND, ABLEISM (2019).

[151] HARPUR, *supra* note 150, at 5–6.

[152] *Id.* at 9–12. Having a disability makes an adult 75–89% more likely to be fired. The likelihood of securing a job is reduced by 40% for adults with physical disabilities and even worse for those with mental disabilities. *See* RICHARD BERTHOUD, THE EMPLOYMENT RATES OF DISABLED PEOPLE 298 (2006).

[153] HARPUR, *supra* note 150, at 2, 14–15.

perfectly capable workers being disfavored and disadvantaged. Algorithms satisfying bias preserving metrics can transport these ableist assumptions into future decision-making.[154]

Criminal justice is likewise heavily impacted by racial bigotry, with people of color significantly more likely to be stopped and searched, arrested, and receive harsher punishments than others in the United States[155] and UK.[156] Lending decisions are also plagued with racial bias;[157] race-based "redlining" of postal codes, for example, is still practiced in 2020[158] despite being illegal in many countries.[159] Healthcare is another troubling area. Participant samples in clinical trials and health studies are routinely biased towards white males.[160]

---

154 On ideas on how to counter AI bias in the realm of disabilities, see Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach et al., *Toward Fairness in AI for People with Disabilities: A Research Roadmap* (Aug. 2, 2019) (Position paper, ACM ASSETS 2019), https://arxiv.org/pdf/1907.02227.pdf; Os Keyes, *Automating Autism: Disability, Discourse, and Artificial Intelligence*, 1 J. SOCIOTECHNICAL CRITIQUE 1 (2020).

155 In the United States, African-American people spend on average the same amount in prison for drugs as White people do for violent crimes. African-Americans are incarcerated at a rate eight times higher than that of Whites. In 2009, the majority (70%) of prisoners in the United States were African-American or Latino as opposed to the 1950s, when segregation was still legal, and around 30% of the African-American population was imprisoned. *See* David Cole, *Can Our Shameful Prisons be Reformed?*, GEO. UNIV. L. CTR. (Nov. 19, 2009), https://scholarship.law.georgetown.edu/cgi/viewcontent.cgi?article=1378&context=facpub.

156 Reports from the UK in 2020 show that Black/BAME people are several times more likely to be stopped and searched (9x / 4x) or arrested (3x / 1.5x) than White people. *See* Vikram Dodd, *Black People Nine Times More Likely to Face Stop and Search Than White People*, GUARDIAN (Oct. 27, 2020), http://www.theguardian.com/uk-news/2020/oct/27/black-people-nine-times-more-likely-to-face-stop-and-search-than-white-people. Similar inequality can be found for drug-related arrests and sentencing. *See The Numbers in Black and White: Ethnic Disparities in The Policing and Prosecution of Drug Offences in England and Wales*, RELEASE, https://www.release.org.uk/publications/numbers-black-and-white-ethnic-disparities-policing-and-prosecution-drug-offences (last visited Feb. 21, 2021).

157 For example, in the United States African-Americans, Latinx, and immigrants are the main targets for predatory lenders which contributed to the 2008 housing crisis. *See* STIGLITZ, *supra* note 7, at 88.

158 Patrick Rucker, *Trump Financial Regulator Quietly Shelved Discrimination Probes into Bank of America and Other Lenders*, PROPUBLICA, https://www.propublica.org/article/trump-financial-regulator-quietly-shelved-discrimination-probes-into-bank-of-america-and-other-lenders?token=nD-X136_tDm0nh1l4Xtv0LbpjY_BSO3u (last visited Feb. 28, 2021).

159 In Germany for example, it would have been illegal to solely use postcodes for lending decisions. *See* Wachter & Mittelstadt, *supra* note 7, at 589–90. For U.S. history on the topic, see Willy E. Rice, *Race, Gender, Redlining, and the Discriminatory Access to Loans, Credit, and Insurance: An Historical and Empirical Analysis of Consumers Who Sued Lenders and Insurers in Federal and State Courts, 1950–1995*, 33 SAN DIEGO L. REV. 583 (1996); *see also* HALLEY ET AL., *supra* note 12, at 110.

160 PEREZ, *supra* note 17, at 115–16. On how to address bias in the medical setting, see Timo Minssen, Sara Gerke, Mateo Aboy, Nicholson Price et al., *Regulatory Responses to Medical Machine Learning*, J.L. & BIOSCIENCES, APR. 11, 2020, at 18, and Mirjam Pot, Wanda Spahl &

Medical interventions and metrics often fail to account for biological differences between populations,[161] for example by treating women simply as "little men" despite health conditions often presenting differently in men and women.[162] The existence of "neutral" health data concerning sexual identity is particularly difficult to imagine given the clinical designation of "homosexuality" as a mental illness until the 1970s. Likewise, being transgender was formally classified as a type of identity disorder until 2012.[163]

Certain people thus face systematic disadvantages in the workplace, lending, education, criminal justice, health, insurance, and other areas. This is morally and legally problematic. Merit alone is often insufficient for individuals to succeed compared to their peers.[164] Treating the status quo as neutral does not sufficiently acknowledge this social reality. In other words, by merely seeking to preserve per-group error rates, bias preserving methods implicitly overestimate the role of meritocracy.[165] This can be problematic; in Western societies, factors such as inheritance, luck, unequal opportunity, and discrimination are just as important to success as merit.[166] For example, the best

---

Barbara Prainsack, *The Gender of Biomedical Data: Challenges for Personalised and Precision Medicine*, 9 SOMATECHNICS 170 (2019).

[161] PEREZ, *supra* note 17, at 116. One of the reasons why this is not done is because it is more complex (e.g., fluctuating hormone levels during the menstrual cycle), risky (e.g., female participants could be pregnant), and time and resource intensive to study women. SAINI, *supra* note 13, at 58.

[162] Women differ for example in percentage of body fat, skin thickness, hormone levels and compositions, changing hormone levels throughout the menstrual cycle, and changing hormone levels prior to puberty and after menopause. Each of these factors affect how well drugs work or how much we are affected by toxins or environmental impacts. *See* PEREZ, *supra* note 17, at 116; SAINI, *supra* note 13, at 59, 62.

[163] HALLEY ET AL., *supra* note 12, at 58. Specifically, until the 1970s the Diagnostic and Statistical Manual of Mental Disorders ("DSM") classified "homosexuality" as a mental illness. *Id.*

[164] STEPHEN J. MCNAMEE & ROBERT K. MILLER, THE MERITOCRACY MYTH 243 (2009).

[165] Judith Butler refers to it as the bootstrapping argument. JUDITH BUTLER, GENDER TROUBLE: FEMINISM AND THE SUBVERSION OF IDENTITY (2011). According to the myth, a system can be considered fair if all people have the same opportunities to succeed (i.e., formal equality, equality of treatment). Success or failure rests solely on the merit of the individual. The possibility of institutional inequality disadvantaging certain (groups of) people is discounted. *Id.*

[166] Specifically, meritocracy is a myth because "of the combined effects of non-merit factors such as inheritance, social and cultural advantages, unequal educational opportunity, luck and the changing structure of job opportunities, the decline of self-employment, and discrimination in all of its forms." Stephen J. McNamee & Robert K. Miller, Jr., *The Meritocracy Myth*, SOCIATION TODAY, http://www.ncsociology.org/sociationtoday/v21/merit.htm (last visited Feb. 28, 2021). Also, evidence from the 2011 Economic Mobility Project shows a strong link between parental education and children's economic, educational, and socio-motional outcomes in many countries, and most strongly in the United States. *Does America Promote Mobility As Well As Other Nations?*, ECON. MOBILITY PROJECT, https://www.pewtrusts.org/-/media/legacy/uploadedfiles/pcs_assets/2011/critafinal1pdf.pdf (last visited Feb. 28, 2021). In

predictor for whether a person will be in poverty as an adult is whether they were born into poverty.[167]

Of course, non-discrimination law has helped to remedy the effects of many social biases and inequalities. In certain areas such as the workplace or when offering goods and services (both of which increasingly face greater deployment of AI and automated decision-making), the law protects certain groups (e.g., gender, race, sexual orientation, disability) from direct discrimination and indirect discrimination.[168]

Unfortunately, changes in law do not equate directly to changes in mindsets. Many racist and sexist practices historically seen as "justified" still remain in practice and legacy. Granting women access to managerial jobs, for example, does not fix inequality in employment overnight; rather, substantive equality is only possible at a societal level when fair practices and rules have been in place for multiple generations.[169]

The data we use to train models and make automated decisions carries the legacy of our unequal past and present. By treating the status quo as neutral, bias preserving metrics miss out on the opportunity to shed light on, and begin

---

general and citing this work, see STIGLITZ, *supra* note 7, at 22. The countries surveyed were the United States, UK, France, Germany, Sweden, Italy, Australia, Finland, Denmark, and Canada. *Id.*

167 EUBANKS, *supra* note 20, at 205. In the United States, for example, 40% of children from the poorest income group remain poor. Those who move up tend to only move up a little as adults, whereas individuals born into the highest income group tend to remain there. *See* Markus Jantti, *American Exceptionalism in a New Light: A Comparison of Intergenerational Earnings Mobility in the Nordic Countries, the United Kingdom and the United States* 17 (Jan. 2006), https://www.econstor.eu/bitstream/10419/33604/1/513429301.pdf. In Nordic countries, on the other hand, only 20% remain in the poorest group as adults, 30% in the U.K. *See* HALLEY ET AL., *supra* note 12, at 106; STIGLITZ, *supra* note 7, at 23. Similarly, "differences in initial conditions as of a real-life age of 23 account for more of the variation in realized lifetime earnings, lifetime wealth, and lifetime utility than do shocks over the working lifetime." Mark Huggett, Gustavo Ventura & Amir Yaron, *Sources of Lifetime Inequality*, 101 AM. ECON. REV. 2923, 2949 (2011). *See also* Alan B. Krueger, *The Rise and Consequences of Inequality in the United States*, 12 SPEECH AT THE CTR. FOR AM. PROGRESS 3 (2012) (explaining "[t]he chance of a person who was born to a family in the bottom 10% of the income distribution rising to the top 10% as an adult is about the same as the chance that a dad who is 5'6" tall having a son who grows up to be over 6'1" tall. It happens, but not often"). Similarly, poor children who succeed academically have been found to be less likely to graduate from college than richer children who did worse in school and tend to remain worse off comparatively. *Id. See also* STIGLITZ, *supra* note 7, at 24; Jonathan Chait, *The Ideological Fantasies of Inequality Deniers*, N.Y. MAG.: INTELLIGENCER (Oct. 26, 2011), https://nymag.com/intelligencer/2011/10/the_ideological_fantasies_of_i.html. Further, a 2020 OECD study on social mobility shows that it takes six generations in Germany to move from the lowest income bracket to an average salary, five generations for the United States, Switzerland, Austria, and at least three generations in the Finland and Sweden. *See* Schneller Schlau, *Wenn die Eltern nicht studiert haben*, WIRTSCHAFT, https://www.faz.net/aktuell/wirtschaft/schneller-schlau/sozialer-aufstieg-wenn-die-eltern-nicht-studiert-haben-16960036.html (last visited Feb. 28, 2021).

168 Wachter, *Affinity Profiling*, *supra* note 42; Wachter et al., *supra* note 34.

169 *See supra* Part II.

to address, the systemic causes of inequality. To combat systemic inequality and achieve substantive equality, private and public actors must play an active role.[170] In this context, choosing to preserve the status quo must be treated as an explicit normative decision that deems the status quo acceptable.[171] If this choice occurs in sectors known to be marked by injustice, it can potentially be seen as conflicting with the substantive aims of the law. Such a choice raises prima facie discrimination and ought to be justified.

To assess and potentially justify prima facie discrimination in fair machine learning, it is essential to recognise the diverse manifestation of inequalities that occur globally. Gender and racial discrimination and other issues of bigotry in the United States and the Member States of the EU will manifest differently according to the cultural and historical legacies of individual countries. One type of inequality cannot be easily be assumed to also occur in a different environment. What we have coined "contextual equality" must factor into the choice and justification of fairness metrics and inherited biases in machine learning and AI.[172]

V. TOWARDS SUBSTANTIVE EQUALITY IN FAIR MACHINE LEARNING

As demonstrated above, the status quo is marked by significant implicit and explicit bias and inequality. Using past decisions as a basis for future automated decisions means past biases can easily be inherited by a trained model.[173]

Returning to the context of indirect discrimination occuring in our hypothetical scenario in court,[174] we argue that using bias preserving metrics in contexts where unjustified bias and inequality have existed historically can give rise to prima facie discrimination. As mentioned above,[175] under normal circumstances a claimant would need to provide evidence to convince the court that prima facie discrimination exists by way of showing that an "apparently neutral provision, criterion or practice" disproportionately disadvantages a protected group in comparison with other people.

---

170 *See supra* Section II.C.

171 EUBANKS, *supra* note 20; O'NEIL, *supra* note 20.

172 Elsewhere, we have coined the term "contextual equality" to describe the contextual application of non-discrimination law by the judiciary in the EU. Examination of relevant jurisprudence reveals that fairness and discrimination are fluid concepts that are given meaning on a case-by-casis basis. For the argument for contextual equality in full, and its significance for fair machine learning, see Wachter et al., *supra* note 34.

173 The only case in which bias is not inherited is a hypothetical utopia in which past decisions are perfectly fair or in which all people receive the minimal possible outcome ("levelling down").

174 *See supra* Section III.A.

175 *See supra* Section II.A.

By definition, high accuracy models trained on historical data to satisfy a bias preserving metric will often replicate the bias present in their training data. This feature makes the hypothetical claimant's task of establishing prima facie discrimination simpler. The claimant will not need to gather substantial evidence demonstrating the disparate nature of the contested "provision, criterion or practice" itself. Rather, the claimant need only show that significant disparity or bias has historically existed in the decision-making context (e.g., in employment) to prove that the contested "provision, criterion or practice" (e.g., an automated decision-making model trained with a bias preserving metric[176]) is prima facie discriminatory.

Once prima facie discrimination has been established, the burden of proof shifts to the alleged offender (e.g., the actor using an automated decision-making system) who then must justify the contested "provision, criterion or practice" under the proportionality test citing a legitimate interested. Contested measures can be justified if there is a legitimate interest, and the means to achieve it are necessary and proportionate.[177] Given the ease of establishing prima facie discrimination, using bias preserving metrics as a basis for automated decision-making should be accompanied by consideration of possible justifications far in advance of actual litigation.

But a key challenge remains for justifying bias preserving metrics for decision-making purposes. The proportionality test states that for a contested rule or practice to be classified as "necessary," there must be no other less infringing means to achieve the interest in question. Bias transforming metrics can be seen as "less infringing means" because they are better suited to promoting substantive equality. Specifically, unlike bias preserving metrics, they give the decision-maker a choice of the properties a classifier should exhibit. They do so through the choice of conditioning variable(s) for conditional independence or the choice of metric for fairness through awareness.[178] In doing so, they allow for a less intrusive or biased metric to be selected as the basis for decisions. As a result, arguments offered to justify usage of bias preserving metrics might fail because they may not be considered "necessary" in a legal sense unless it can be shown that conditioning on the target variable is the least intrusive possible means of achieving a legitimate interest.[179]

---

176 Note that typical machine learning systems trained without any form of fairness constraint also look to replicate the past decisions made on historic data with high fidelity, and the same argument can be made regarding them.

177 For an overview of the ECJ's case law on legally accepted justifications, see Wachter, *Data Protection*, *supra* note 3, at 46–54; Wachter et al., *supra* note 34, at 534–37.

178 Cynthia Dwork, Mortiz Hardt, Toniann Pitassi, Omer Reingold et al., *Fairness Through Awareness* (Nov. 30, 2011), https://arxiv.org/pdf/1104.3913.pdf.

179 Justification would only be possible if the alleged offender could show that bias preserving metrics are less infringing in a legal sense than bias transforming metrics. It is difficult to imagine a scenario in the context of EU non-discrimination law where this would be the case;

This is not to suggest that using bias transforming metrics in automated decision-making will eliminate historical biases or prevent future disparity altogether. Their usage likewise does not establish a legal duty to dismantle inequality.[180] They cannot force decision-makers to change their behaviors or criteria. The same requirements apply to contested practices using these metrics; any disparity found must still be justified as necessary, proportionate, and in pursuit of a legitimate interest.

Bias transforming metrics are not a "silver bullet" to solve algorithmic discrimination. Rather, their value comes from their explicit requirement that users must make a normative judgement of what bias is acceptable in a given use case.[181] This is, of course, a politically, legally, and ethically significant decision in itself. Nonetheless, bias transforming metrics force designers and decision-makers to confront fairness and to consider the biases and inequalities in their data that would otherwise be ignored, hidden, or treated as justified by bias preserving metrics.[182]

When viewed as a tool to confront past disparity, bias transforming metrics have two clear benefits. First, in the context of litigation, bias transforming metrics force otherwise ignored inequalities into the conversation. Decision-makers must then explain why the disparity is justified, or why it should be ignored. Open and transparent discussion of the justifiability of disparity in this manner is essential to promote substantive equality. Second, discussing disparity and the relative intrusiveness of possible fairness metrics creates an opportunity. It gives decision-makers or developers of automated systems a chance to tweak the decision-making process and criteria to level the playing field for disadvantaged groups, unless a legal justification for the disparity can be demonstrated. When used correctly, bias transforming metrics help ensure inequality in automated decision-making is explicitly acknowledged, discussed, and potentially justified in a consistent and realistic manner. This type of critical self-reflection and foresight is essential if users of AI, machine learning, and automated decision-making in the public and private sectors are to play a more active role in dismantling inequality.[183]

---

hypothetically, formal equality may be valued higher in certain decision-making contexts in which case bias preserving metrics could be considered necessary.

180 *See supra* Section II.C.

181 For example, in the case of conditional independence, the choice of bias is made by choosing which variable(s) to condition on.

182 Certain bias transforming metrics can, when coupled with summary statistics, help identify hidden inequalities by treating all groups as equal and report on differences in outcomes between them. *See infra* Section VI.B.

183 For a discussion of effective usage of bias transforming metrics, specifically Conditional Demographic Disparity, see Wachter et al., *supra* note 34, at 547.

While the existence and precise requirements of positive duties remains debated,[184] critically investigating historical bias in this manner can only benefit designers and users of automated decision-making. In EU non-discrimination law, intent is not necessary to establish direct or indirect discrimination.[185] In practice, this means decision-makers have an interest to test their procedures as thoroughly as possible because they can be held liable for disparity independently of their prior knowledge. A lack of intent is not an effective justification in court. In fact, even well-intentioned actions can be seen as discriminatory.[186] This again means careful thought needs to occur before bias preserving metrics are deployed as decision-making criteria where an obligation to promote substantive equality exists.[187]

For all these reasons, unquestioning use of bias preserving metrics in automated decision-making is therefore inadvisable in places governed by non-discrimination law and related legal frameworks that aim at substantive equality, such as the UK and EU. To actively move towards substantive equality in fair machine learning, we recommend usage of bias transforming metrics for purposes of decision-making.

With that said, bias preserving metrics still have a role to play in fair machine learning. In legal contexts that pursue formal rather than substantive equality, or for use cases where existing biases are normatively acceptable, bias preserving metrics may be preferable. For purely diagnostic and testing purposes (i.e., not decision-making), both bias preserving and transforming metrics are broadly acceptable. Ideally, users should test as broadly as possible with both bias preserving and transforming metrics to investigate the fairness of their decision-making systems.

---

184 *See supra* Section II.C.

185 EUROPEAN UNION AGENCY FOR FUNDAMENTAL RTS. AND COUNCIL OF EUR., HANDBOOK ON EUROPEAN NON-DISCRIMINATION LAW 239 (2018 ed.), https://fra.europa.eu/sites/default/files/fra_uploads/1510-fra-case-law-handbook_en.pdf.

186 *Id.* at 240.

187 *See supra* Section II.C. In American non-discrimination law, intent is relevant to establish whether disparate treatment (direct discrimination) has occurred. Testing for discrimination plays a different role in this context, as it can potentially reveal previously unknown inequality to the decision-maker. Failure to act to correct this inequality could then potentially be seen as evidence of intent to commit discrimination. Investigating the connection between fairness testing and intent is an interesting future question for work on UK, EU, and U.S. non-discrimination law and automated decision-making. For more details on these doctrines, see Nachbar, *supra* note 65, at 23–24. *See also id.* at 51, 55 (explaining that if knowledge of the discriminatory nature of a "facially neutral" practice exists, it could turn disparate impact into (intentional) disparate treatment).

VI. CONCLUSION AND RECOMMENDATIONS

As a field, fair machine learning is predominantly driven by statistical measures of fairness and fixes that address "technical bias." This approach ignores important, explicit normative decisions about how a system should behave and risks leaving important legal, ethical, and political decisions solely to developers, deployers, and users. These decisions determine what is fair and discriminatory, whether a "particular disadvantage" was severe enough to warrant discussion, and ultimately whether indirect discrimination can be justified.[188]

In this Article we introduced a new classification scheme for fairness metrics to clarify the lines of debate and make clear the normative and political dimensions of technical work on fair machine learning. Put simply, developers have a choice between two types of metrics: (1) "bias preserving" metrics that take society as it currently exists as a neutral starting point or "level playing field" from which we can measure inequality and bias in machine learning; and (2) "bias transforming" metrics that acknowledge historical inequalities and start from the assumption that certain groups will have a worse starting point than others.

While technical fixes alone cannot solve the root causes of societal inequalities, our choice of fairness metric can ensure machine learning applications do not exacerbate existing inequalities and fully acknowledge the extent and significance of existing inequalities. The choice of variables to condition on for fairness tests, thresholds for illegal disparity, and acceptable arguments to justify disparity are difficult political determinations.

Ultimately, these determinations will be made by a court, subsequent case law, and potentially even new laws. However, using bias transforming metrics draws further attention to these important determinations and helps ensure they are made in the open involving democratically legitimised courts and legislatures. To advance the adoption of bias transforming metrics in practice, we conclude with several practical and policy recommendations and open questions for future research.

A. A Checklist for Choosing Appropriate Fairness Metrics

We have argued that bias preserving metrics in decision making can give raise to prima facie indirect discrimination under EU non-discrimination law.[189] Developers should proactively justify the potentially discriminatory effect of their "provision, criterion or practice" under indirect discrimination doctrine by providing an objective justification under the proportionality test (i.e., a

---

188 For more detail, see Wachter, *Data Protection*, *supra* note 3; Wachter et al., *supra* note 34.

189 *See supra* Part V.

legitimate interest that is pursued in a necessary and proportionate manner).[190] This need for legal justification reflects our observation that the usage of fairness metrics is not a neutral choice. It is an explicit normative decision, and must be treated as such.

To assist in this process of choosing appropriate fairness metrics for both diagnostic and decision-making purposes in machine learning, Figure 1 presents a checklist reflecting the contributions and recommendations made throughout this Article. This simple checklist is intended for use by developers, deployers, and other users of AI, ML, and automated decision-making systems.

Question 1 reflects the distinction between using fairness metrics to test for and diagnose disparity, and to make fair decisions in practice. Both bias preserving and transforming metrics are valuable for diagnostic purposes.[191] Substantive decisions are those with impact on individuals falling within the remit of non-discrimination law.

---

190 For an overview of the ECJ's case law on legally accepted justifications, see Wachter, *Data Protection*, *supra* note 3, at 46–54; Wachter et al., *supra* note 34, at 534–37.

191 *See supra* Section III.C.

Question 2 addresses the need for justification when using bias preserving metrics to make substantive decisions in contexts historically marked by inequality. Inequality is widespread in society.[192] Following recommendations from legal scholars, we advise developers, deployers, and users to reverse the burden of proof by taking for granted the existence of inequality unless explicitly disproven or justified. Question 2 should therefore only be answered in the negative where historical inequality can be shown not to exist in the given decision-making context, or existing inequality has already

Figure 1: Bias preservation checklist

Q1: Are you using fairness metrics to solely diagnose disparity, but are not making substantive decisions about individuals?

Yes: Both bias preserving and transforming metrics can be used.
No: Go to Question 2.

Q2: Are you deploying a system to make decisions in an area known to have unacceptable historical social inequality?

Yes: Go to Question 3.
No: Recommend investigation of possible bias in use case before choosing a metric.

Q3: Are you deploying the system and in a legal jurisdiction that **solely** promotes formal equality?

Yes: Both bias preserving and transforming metrics can be used.
No: Go to Question 4.

Q4: Are you deploying the system and in a legal jurisdiction that promotes substantive equality?

Yes: Recommend using a bias transforming metric.
No: Both bias preserving and transforming metrics can be used.

192 *See supra* Part IV.

been deemed legally justified through litigation. If this proves to be the case both bias preserving and transforming metrics can be used.

Questions 3 and 4 distinguish between use cases according to the type of legal framework in place, specifically between those that strictly pursue formal equality, and those aiming at substantive equality.[193] The legal acceptability of fairness metrics varies according to the emphasis local legal frameworks place on formal or substantive equality. Bias transforming metrics are best placed for (automated) decision-making aimed at substantive equality. If Question 3 is answered in the affirmative, meaning a system is being used within a framework solely aiming at formal equality, both bias preserving and transforming metrics can be used to pursue this aim.

In contrast, if Question 4 is answered in the affirmative, meaning the framework at hand aims at substantive equality, we recommend usage of only bias transforming metrics for decision-making purposes in automated systems. This recommendation follows the capacity of bias transforming metrics to facilitate dialogue around the existence and justifiability of social bias and inequality, and to give developers, deployers, and users a choice of the bias the system should exhibit. This choice of, for example, variables to condition on (in the case of conditional independence) creates a clear path to open dialogue about the legal acceptability of existing disparity.

Bias preserving metrics are less well-suited to this purpose but can, of course, still be used for decision-making in the context of substantive equality. However, if used, we recommend developers, deployers, and users pre-emptively consider how to justify bias inherited by a system due to the choice fairness metric. Recognizing the possibility of future litigation, this justification should follow the model set by indirect discrimination and the proportionality test because bias preserving metrics can easily give rise to prima facie discrimination.[194]

B. Using Bias Transforming Metrics To Support Substantive Equality

We have argued elsewhere that CDD, a type of conditional independence and bias transforming metric, is the fairness metric most compatible with the concepts of equality and illegal disparity as developed by the European Court of Justice.[195] This compatibility lends increased legal legitimacy to the usage of the metric by public and private actors to measure bias and fairness in AI and algorithmic decision-making systems.

---

193 Determining the type of legal framework at hand is typically a question of politics and the application and interpretation of the law, and may change over time. As we have suggested above, EU non-discrimination law aims at substantive equality. *See supra* Section II.B.

194 *See supra* Section III.C; Part V.

195 Wachter et al., *supra* note 34.

CDD treats all people (groups) as equal, meaning they should be treated the same. The test flags up any disparity between groups that remains once an appropriate conditioning variable has been applied. This notion of fairness follows the Aristotelian postulate of treating "like cases alike" and enables formal equality.

At the same time, CDD enables substantive equality by flagging up for further discussion any relative disparity between groups in a given population over a set of decisions or other outcomes. Often this disparity will be subtle, unexpected, or systemic, but likewise unjustified and requiring correction in the decision procedure. These findings can be published as summary statistics and function as an early alarm system for potentially illegal disparity in automated decision-making.[196]

CDD of course has limitations and is not a silver bullet for algorithmic fairness.[197] Choosing the right conditioning variables is a political decision and developers can be inclined to choose favorable conditions. However, if these choices are—as we recommend—published as summary statistics, these conditions are transparent and are open to inspection and rebuttal.[198]

CDD and other bias transforming metrics thus enable public and private actors to take a more active role in establishing substantive equality. They can spark dialogue between developers, claimants, regulators, and courts to determine their respective roles, duties, and obligations to realise substantive equality. Where unjustified disparity is identified, processes may need to be adapted, for example by changing decisions criteria, adding different variables, or giving different weight to existing ones (e.g., telling a model to give less importance to salary or career breaks because they are gender biased proxies for job performance). This can help create decision criteria that better measure merit.

This potential usage of CDD (and other bias transforming metrics) to promote substantive equality should not be confused with a requirement for positive action (affirmative action), for example a requirement to hire people because of their gender.[199] On the contrary, bias transforming metrics can help identify talented job applicants that are undervalued by biased decision criteria that fail to consistently and fairly reflect merit and competence across all job applicants.[200]

---

196 *Id.*

197 *See supra* Section III.C.

198 Wachter et al., *supra* note 34, at 555–57.

199 *See supra* Section II.C.

200 Statistics show, for example, that people who have a criminal record are more diligent and dedicated workers in certain sectors (e.g., customer service). *See* Burdon & Harpur, *supra* note 4, at 688. Despite this, hiring criteria that reject applicants with criminal records are the norm. Using a bias transforming metric in this context could simultaneously promote substantive equality and help companies hire more reliable workers.

C. *Substantive Equality Duties in Fair Machine Learning*

Whilst it is clear that the rationale and the aim of non-discrimination law is to dismantle inequalities,[201] it is an open question as to what is expected of different (private and public) stakeholders. Legal and policy scholars continue to debate the existence and specific requirements for proactive, positive duties under non-discrimination law for both the public and private sector.[202] Specifying the requirements of positive equality duties in fair machine learning is an important area for future research and policy-making.

Positive equality duties can be exercised through the usage of bias transforming metrics. Specifically, dialogue concerning which biases a system should adopt, which variables to condition on (in the case of conditional independence), and which forms of inequality can be justified is key to promote substantive equality in practice. The use of bias transforming metrics to identify and question existing and emergent disparity can ensure this dialogue occurs and includes the right stakeholders; such political determinations should not be made in isolation by developers, deployers, and users of automated systems. Rather, a broad and open dialogue is needed to answer questions such as

- Should income be used as a variable to decide if somebody is granted a loan, given what we know about income equality? What would be a less discriminatory, but equally useful alternative?
- Should the possession of a higher education degree be allowed as a mandatory requirement for employment given the significant inequality in access to education?
- Should we rely on recommendation letters and Grade Point Average even if they have been shown to be biased towards gender and ethnicity?
- Do socially acceptable or even desirable disparities exist? What type of disparities, if any, should be maintained in society?

These are not questions that can be answered by a choice of fairness metric directly, but bias transforming metrics and public summary statistics can spark and inform this crucial dialogue. By using a bias transforming metirc and publishing summary statistics, developers, deployers, and users are forced to make and openly justify their normative judgement as to which biases are locally acceptable. Ultimately courts and legislatures will have to decide. They will need to develop case law and legislation that clarifies the (active or passive) obligation(s) to dismantle inequality and promote substantive equality that should exist for different stakeholders. They will likewise need to clarify which justifications for disparity should be accepted in practice, and across different use cases.

---

201 De Vos, *supra* note 56.

202 *See supra* Section II.C. Examples include public sector equality duties in the UK, and evidential requirements to successfully justify disparity in the context of indirect discrimination. *See* FREDMAN, *supra* note 50.

While bias transforming metrics cannot directly provide a definitive answer regarding the existence and requirements for positive equality duties, their usage creates opportunities for incremental change. This is an important shift away from using machine learning, intentionally or otherwise, to further entrench the status quo with bias preserving metrics. Bias transforming metrics and summary statistics can be seen as a roadmap for societal change in the workplace, lending, education, criminal justice, health, insurance, and other areas.

D. More Data Alone Is Not the Answer

Finally, bias often occurs not for any technical reasons, but rather because a dataset is not representative of the population. Many critical data gaps exist due to limitations on resources, access, or motivation. In healthcare, gaps in data for female and BAME patients are unlikely to be closed soon. The same holds true for missing data on real events, such as cases of (sexual) harassment, hate crimes or violence against women, people of color, or LGBTQ people. Cases often go unreported due to a lack of reporting mechanisms, weak legal protection, and the low conviction rates of cases brought forward.[203]

Inequality has many and diverse faces. One form of bigotry cannot be assumed to also exist or manifest in the same way elsewhere. More data is required to investigate multi-faceted inequality internationally to promote what we have elsewhere called "contextual equality."[204]

Recognising this, these gaps in awareness of inequality can motivate more extensive collection of data about protected groups. It is a generally accepted fact that in order to prevent discriminatory or biased outcomes, data about protected groups must be collected.[205] Failure to collect this data will not prevent discrimination against protected groups, but perhaps make it more difficult to detect.[206] Sensitive data is needed to test whether automated decision-making discriminated against groups based on protected attributes (e.g., data on race, disability, sexual orientation).[207]

---

203 *See supra* Part IV.

204 For more on "contextual equality," see Wachter et al., *supra* note 34.

205 On the practical limitations of data collection in the EU, see TIMO MAKKONEN, MEASURING DISCRIMINATION DATA COLLECTION AND EU EQUALITY LAW (2007).

206 DWORK ET AL., *supra* note 130; KUSNER ET AL., *supra* note 130; Wachter et al., *supra* note 34, at 527–28; Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35 (2013); Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardsiel et al., Proxy Non-Discrimination in Data-Driven Systems (July 25, 2017) (Companion paper), https://arxiv.org/pdf/1707.08120.pdf.

207 KUSNER ET AL., *supra* note 130; Chris Russell, Matt J. Kusner, Joshua Loftus, Ricardo Silva, *When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 6396–405 (2017).

Naturally, privacy scholars have urged to be mindful of the privacy implications of such privacy invasive data collection.[208] This is a legitimate concern and closely related to troubling historical experiences that have significantly harmed minority and marginalized groups in society.

The collection and evaluation of data is seen as a product of the Enlightenment. Decision-making that is based on "ground-truth" and derived from scientific methods rather than just religious dogma was seen as a step forward.[209] But scientific research, data collection, and databases also contributed to eugenics in Europe, the UK[210] and the United States,[211] genocide during WWII, racist immigration practices and the denial of basic human rights in the United States,[212] justification of slavery,[213] forced sterilisation in the UK,[214] United States, Germany and Puerto Rico from the early to the mid 20th Century,[215] punishment, castration and imprisonment of LGBT members,[216] and denial of equal rights and protection (e.g., against sexual violence) to women.[217]

---

208 VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA : A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK AND THINK (2013). For a United States and EU comparison, see Joris Van Hoboken, *From Collection To Use in Privacy Regulation? A Forward Looking Comparison of European and US Frameworks for Personal Data Processing*, 231 EXPLORING THE BOUNDARIES OF BIG DATA (2016). For an international view, see 63 LEE A. BYGRAVE, DATA PRIVACY LAW: AN INTERNATIONAL PERSPECTIVE (2014). For an European view, see Sandra Wachter, *Normative Challenges of Identification in the Internet of Things: Privacy, Profiling, Discrimination, and the GDPR*, 34 COMPUT. L. & SEC. REV. 436–49 (2018); Sandra Wachter, *The GDPR and the Internet of Things: a Three-step Transparency Model*, 10 L. INNOVATION & TECH. 266–94 (2018). For a EU and German view, see Mario Martini, Wiebke Fröhlich & Saskia Fritzsche, *Algorithmen als Herausforderung fu r die Rechtsordnung* (2017). For empirical evidence of mobile data collection, see Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao et al., *Third Party Tracking in the Mobile Ecosystem*, *in* PROCEEDINGS OF THE 10TH ACM CONFERENCE ON WEB SCIENCE 23–31 (2018). On online harms, see Woods Lorna & Perrin William, *An Updated Proposal by Professor Lorna Woods and William Perrin*, https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/01/29121025/Internet-Harm-Reduction-final.pdf (last visited Feb. 28, 2021).

209 HALLEY ET AL., *supra* note 12, at 9.

210 This happened until the 1930s. *See* EDDO-LODGE, *supra* note 15, at 20–21.

211 HALLEY ET AL., *supra* note 12, at 36.

212 *Id.* at 25.

213 *Id.* at 36–37.

214 EDDO-LODGE, *supra* note 15, at 20–21.

215 HALLEY ET AL., *supra* note 12, at 36–38.

216 *Id.* at 15–17.

217 SAINI, *supra* note 13, at 233–35.

Privacy[218] and harms from slippery slopes in data collection[219] are legitimate concerns and must be taken seriously.

Setting these concerns aside for a moment, a more fundamental challenge must be addressed. One could be tempted to think that problems of bias and fairness in machine learning will be naturally solved by collecting more (sensitive) data and closing gaps in representation. However, it is naïve to assume that fair and equal outcomes will necessarily result from more data being collected.

Awareness of inequalities is not the same as rectifying them.[220] Pay gaps based on gender and race are painful examples of this reality.[221] In 2016 in the UK, for example, there was an 18 to 23% gap in wages between men and women (depending on the sector).[222] Gender discrepancies are nothing new, and yet extensive knowledge of them has not yet led to their elimination around the world.[223]

Their persistence suggests that significant political, social, and legal effort is needed to overcome well-established social and economic inequality. These are longstanding challenges that cannot be solved through technological fixes or by simply choosing the right metric to measure fairness in machine learning. Rather, open and collaborative dialogue involving computer scientists and developers, lawyers, ethicists, social scientists, regulators, the general public, and many others is essential.

Countering inequalities requires intentional and often cost-intensive changes to decision processes, business models, and policies. To justify further collection and usage of sensitive data, it is necessary to first demonstrate serious commitment and political will to rectifying inequality.

A first step towards demonstrating this commitment in practice is through proactive fulfilment of positive duties around substantive equality. Choosing to maintain the status quo by using bias preserving fairness metrics

---

218 On how data will follow us forever, see VIKTOR MAYER-SCHÖNBERGER, DELETE (2011) (ebook).

219 For a discussion of surveillance and chilling effects, see Jon Penney, *Chilling Effects: Online Surveillance and Wikipedia Use*, 31 BERKELEY TECH. L.J. 117 (2016).

220 EDDO-LODGE, *supra* note 15, at 208.

221 Fredman, *supra* note 86, at 6.

222 *See* SAINI, *supra* note 13, at 6–7 (citing statistics from 2016).

223 The global gender pay gap varies greatly from country to country. A World Economic Forum study published in 2020 shows how countries around the world have closed their gaps since 2006. Western Europe has been the best performing, with countries such as Iceland (87.7), Norway (84.2), Finland (83.2) and Sweden (82.0) leading the way. WORLD ECONOMIC FORUM, GLOBAL GENDER GAP REPORT 2020, http://www3.weforum.org/docs/WEF_GGGR_2020.pdf. However, in terms of economic participation and opportunity Western Europe lags behind (69.3%) other players such as North Africa (75.6%), Eastern Europe, and Central Asia (72.2%). *Id.* North America has closed 73% of its pay gap, sub-Saharan Africa has closed 68% and South Asia two thirds. *Id.*; *see also* PEREZ, *supra* note 17, at 75–78.

cannot be considered a neutral choice in this regard; rather, it must be understood as a legally significant choice requiring explicit consideration by AI developers, users, and regulators going forward.

Appendix 1 – Table of Fairness Metrics

Table 1a shows whether standard definitions of algorithmic fairness are bias preserving and satisfied by a perfect classifier. The definitions of fairness considered are those from a 2018 "state of the art" survey paper by Verma and Rubin.[224]

In practice, machine learning practitioners do not simply look for a system that (approximately) satisfies a particular fairness definition—as many fairness definitions can be satisfied by constant classifiers. Instead, they look for a classifier that is as accurate as possible, while still satisfying the fairness metric. As such perfect classifiers that satisfy $\hat{y} = y$ and are 100% accurate are an important case to consider, as this represents the ideal behavior of a classifier.

In the equations below we use $\hat{y}$ for the classifier response y for the target value of the original data. Capital letters represent particular variables with A being the protected attribute that indicates membership of a protected group, e.g., gender or racial. C in definition 2 is a confounding variable that must be explicitly selected. Where the definition of fairness makes use of the inputs to the classifier, we write $x_1, x_2, \ldots, x_n$ for all inputs excluding the protected attribute a. Although the substitution $\hat{y} = y$ transforms $\hat{y}$ in to a discrete classifier, it still satisfies the continuous definitions (i.e., definitions 10-13).

Evaluating if a method is bias-preserving is straightforward. We simply substitute the classifier response $\hat{y}$ with y, and observe if the formula is trivially true.

---

224 Verma & Rubin, *supra* note 128.

Fairness metrics	Formula	Bias preserving?
1. Group fairness, Statistical (demographic) parity[i]	$P(\hat{Y}=1\|A=a) = P(\hat{Y}=1\|A=a')\ \forall a,a'$	✗
2. Conditional statistical (demographic) parity, Conditional independence[ii]	$P(\hat{Y}=1\|C=c,A=a) = P(\hat{Y}=1\|C=c,A=a')\ \forall c,a,a'$	✗
3. Predictive parity, outcome test[iii]	$P(\hat{Y}=1\|Y=1,A=a) = P(\hat{Y}=1\|Y=1,A=a')\ \forall a,a'$	✓
4. False positive error rate balance[iv]	$P(Y=1\|\hat{Y}=0,A=a) = P(Y=1\|\hat{Y}=0,A=a')\ \forall a,a'$	✓
5. False negative error rate balance,[v] Equal opportunity[vi]	$P(Y=0\|\hat{Y}=1,A=a) = P(Y=0\|\hat{Y}=1,A=a')\ \forall a,a'$ Or the equivalent formula $P(Y=1\|\hat{Y}=1,A=a) = P(Y=1\|\hat{Y}=1,A=a')\ \forall a,a'$	✓
6. Equalized odds[vii]	$P(\hat{Y}=1\|Y=i,A=a) = P(\hat{Y}=1\|Y=i,A=a')\ \forall i\in\{0,1\},a,a'$	✓
7. Conditional use accuracy equality[viii]	$P(\hat{Y}=i\|Y=i,A=a) = P(\hat{Y}=i\|Y=i,A=a')\ \forall i\in\{0,1\},a,a'$	✓
8. Overall accuracy equality[ix]	$P(\hat{Y}=Y\|A=a) = P(\hat{Y}=Y\|A=a')\ \forall i\in\{0,1\},a,a'$	✓
9. Treatment equality[x]	$\frac{P(\hat{Y}=0\wedge Y=1\|A=a)}{P(\hat{Y}=1\wedge 0=1\|A=a)} = \frac{P(\hat{Y}=0\wedge Y=1\|A=a')}{P(\hat{Y}=1\wedge 0=1\|A=a')}\ \forall a,a'$	✓
10. Test-fairness or calibration[xi]	$P(Y=1\|\hat{Y}=t,A=a) = P(Y=1\|\hat{Y}=t,A=a')\ \forall t\in\mathbb{R}\ a,a'$	✓
11. Well-calibration[xii]	$P(Y=i\|\hat{Y}=t,A=a) = P(Y=i\|\hat{Y}=t,A=a')\ \forall i\in\{0,1\},t\in R\ a,a'$	✓
12. Balance for positive class[xiii]	$E(\hat{Y}\|Y=1,A=a) = E(\hat{Y}\|Y=1,A=a')\ \forall a,a'$	✓
13. Balance for negative class[xiv]	$E(\hat{Y}\|Y=0,A=a) = E(\hat{Y}\|Y=0,A=a')\ \forall a,a'$	✓
14. Causal discrimination[xv] (direct discrimination)	$\hat{Y}(x_1,x_2,\ldots,x_n,a) = \hat{Y}(x_1,x_2,\ldots,x_n,a')\ \forall a,a'$	*
15. Fairness through unawareness[xvi]	$\hat{Y}$ if a function of x only and not protected attribute a	*
16. Fairness through awareness[xvii]	The distribution of randomized outcomes is k-Lipschitz with respect to a metric defined over the inputs	✗
17. Counterfactual fairness[xviii]	$\hat{Y}_{A\leftarrow a}(x_1,x_2,\ldots,x_n,a) = \hat{Y}_{A\leftarrow a'}(x_1,x_2,\ldots,x_n,a)$	✗
18. No unresolved discrimination[xix] (causal variant of 2)	$\hat{Y}_{A\leftarrow a,x_k\leftarrow x_k}(x_1,x_2,\ldots,x_n,a) = \hat{Y}_{A\leftarrow a',x_k\leftarrow x_k}(x_1,x_2,\ldots,x_n,a)$	✗
19. No proxy discrimination[xx]	No simple formula	✗
20. Path based causal reasoning[xxi]	No simple formula	✗

Table 1a - Bias preserving fairness metrics (full table)

* Indicates that a perfect classifier satisfying $Y = \hat{Y}$ would always satisfy this definition if perfect predictions can be made without explicitly using the protected attribute such as race or sex.

---

[i] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold et al., Fairness Through Awareness (Nov. 30, 2011), https://arxiv.org/pdf/1104.3913.pdf.

[ii] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goal et al., *Algorithmic Decision Making and the Cost of Fairness*, *in* PROC. OF THE 23RD ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 797 (2017); Faisal Kamiran, Indrė Žliobaitė & Toon Calders, *Quantifying Explainable Discrimination and Removing Illegal Discrimination in Automated Decision Making*, 35 KNOWLEDGE AND INFO. SYS. 613 (2013).

[iii] Camelia Simoiu, Sam Corbett-Davies & Sharad Goel, *The Problem of Infra-Marginality in Outcome Tests for Discrimination*, 11 THE ANNALS OF APPLIED STATISTICS 1193 (2017); Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153 (2017).

[iv] Chouldechova, *supra* note iii; Corbett-Davies et al., *supra* note ii.

[v] Chouldechova, *supra* note iii.

[vi] Moritz Hardt, Eric Price & Nati Srebro, *Equality of Opportunity in Supervised Learning*, *in* ADVANCES IN NEURAL INFO. PROCESSING SYS. 3315 (2016).

[vii] *Id.*

[viii] Richard Berk, Hoda Heirdari, Shahin Jabbari, Michael Kearns et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 SOCIO. METHODS & RSCH. 3 (2021).

[ix] *Id.*

[x] *Id.*

[xi] Chouldechova, *supra* note iii.

[xii] Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores (Nov. 17, 2016), https://arxiv.org/pdf/1609.05807.pdf.

[xiii] *Id.*

[xiv] *Id.*

[xv] Sainyam Galhotra, Yuriy Brun & Alexandra Meliou, *Fairness Testing: Testing Software for Discrimination*, *in* PROC. OF THE 2017 11TH JOINT MEETING OF THE EUR. SOFTWARE ENG'G CONF. AND THE ACM SIGSOFT SYMP. ON THE FOUND. OF SOFTWARE ENG'G 498 (Sept. 4, 2017). Not to be confused with the causal methods 17–20 that make use of structured causal models. *See* JUDEA PEARL, CAUSALITY (2d ed. 2009).

[xvi] Dwork et al., *supra* note i.

[xvii] *Id.*

[xviii] Matt J. Kusner, Joshua Loftus, Chris Russel & Ricardo Silva, Counterfactual Fairness (Mar. 8, 2018) (Conference paper, NIPS 2017), https://arxiv.org/pdf/1703.06856.pdf.

[xix] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt et al., Avoiding Discrimination through Causal Reasoning (Jan. 21, 2018) (Conference paper, NIPS 2017), https://arxiv.org/pdf/1706.02744.pdf.

[xx] *Id.*

[xxi] Razieh Nabi & Ilya Shpitser, *Fair Inference on Outcomes*, 32 *in* PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE (May 29, 2017), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/16683/15898; Silvia Chiappa & Thomas P. S. Gillam, Path-Specific Counterfactual Fairness (Feb. 22, 2018), http://arxiv.org/abs/1802.08139.