

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law

Sandra Wachter¹, Brent Mittelstadt² and Chris Russell³

CONTENTS

Abstract	2
1 Introduction	3
2 Formal and substantive equality in non-discrimination law	12
2.1 Indirect discrimination and substantive equality	15
2.1.1 Positive action and substantive equality	16
2.2 Substantive equality is the aim of EU non-discrimination law	17
2.3 Positive duties and requirements for substantive equality	19
3 Bias preservation in fair machine learning	22
3.1 Fairness metrics and non-discrimination law	23
3.2 Bias preserving and bias transforming fairness metrics	25
3.3 Limits of bias preserving and transforming metrics	29
4 The status quo is not neutral	31
5 Towards substantive equality in fair machine learning	37
6 Conclusion and recommendations	40
6.1 A checklist for choosing appropriate fairness metrics	41
6.2 Using bias transforming metrics to support substantive equality	43
6.3 Substantive equality duties in fair machine learning	44
6.4 More data alone is not the answer	45
Appendix 1 – Table of fairness metrics	49

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

² Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, UK.

³ Amazon Web Services, Inc.

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. Recognising 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 metrics address technical bias but ignore the underlying causes of inequality. In this paper 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 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.

1 INTRODUCTION⁴

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.⁵ 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.⁶ 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.⁷ Prior hiring decisions

⁴ 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 immensity 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. 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, Luminate/Omidyar Group, and the Miami Foundation.

⁵ 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: Ofqual's A-level algorithm: why did it fail to make the grade?, THE GUARDIAN (2020), <http://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels> (last visited Jan 17, 2021).

⁶ How Ofqual failed the algorithm test, UNHERD (2020), <https://unherd.com/2020/08/how-ofqual-failed-the-algorithm-test/> (last visited Jan 9, 2021).

⁷ Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Rethinking Data Protection Law in the Age of Big Data and AI*, 2 COLUMBIA BUSINESS LAW REVIEW (2019), https://cblr.columbia.edu/wp-content/uploads/2019/07/2_2019.2_Wachter-Mittelstadt.pdf (last visited Sep 25, 2018); Sandra Wachter, *Data protection in the age of big data*, 2 NATURE ELECTRONICS 6 (2019).

inform future hiring,⁸ past loan and insurance decisions are the basis for future banking strategy and decisions,⁹ past shopping history will impact future offers and prices,¹⁰ previous tenants look like future tenants,¹¹ and the sentences of past criminals inform risk profiling for potential parolees.¹² Decisions about school admissions are no exception.¹³ 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.

⁸ On issues of AI used in the workplace see Jeremias Prassl & Martin Risak, *Uber, taskrabbit, and co.: Platforms as employers-rethinking the legal analysis of crowdwork*, 37 COMP. LAB. L. & POL'Y J. 619 (2015); JEREMIAS PRASSL, HUMANS AS A SERVICE: THE PROMISE AND PERILS OF WORK IN THE GIG ECONOMY (2018); JEREMIAS ADAMS-PRASSL, What if Your Boss Was an Algorithm? *The Rise of Artificial Intelligence at Work* (2019), <https://papers.ssrn.com/abstract=3661151> (last visited Jan 9, 2021); Amit Datta et al., *Discrimination in Online Personalization: A Multidisciplinary Inquiry.*, 81 FAT 20–34 (2018); Mark Burdon & Paul Harpur, *Re-conceptualising privacy and discrimination in an age of talent analytics*, 37 UNSWLJ 679 (2014).

⁹ On algorithms, bias and credit see 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*, SSRN JOURNAL (2020), <https://www.ssrn.com/abstract=3571266> (last visited Jan 8, 2021); Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, THE UNIVERSITY OF CHICAGO LAW REVIEW 29.

¹⁰ On AI bias when offering goods and services see Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995 (2013); Tal Zarsky, *Privacy and Manipulation in the Digital Age*, 20 THEORETICAL INQUIRES IN LAW 157 (2019); Karen Levy & Solon Barocas, *Designing against Discrimination in Online Markets*, 32 BERKELEY TECH. L.J. 1183 (2017); In view of price discrimination Oren Bar-Gill, *Algorithmic Price Discrimination: When Demand Is a Function of Both Preferences and (Mis) Perceptions*, THE HARVARD 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*, 27 HARVARD BUSINESS REVIEW (2016); Frederik Zuiderveen Borgesius, *Algorithmic Decision-Making, Price Discrimination, and European Non-discrimination Law*, EUROPEAN BUSINESS LAW REVIEW (FORTHCOMING) (2019).

¹¹ On bias in online housing markets see also Joshua Asplund et al., *Auditing Race and Gender Discrimination in Online Housing Markets*, 14 ICWSM 24–35 (2020); on the 2008 housing crisis during which algorithmic tools were used see JOSEPH E. STIGLITZ, *THE PRICE OF INEQUALITY: HOW TODAY'S DIVIDED SOCIETY ENDANGERS OUR FUTURE* 239–242 (2012).

¹² On algorithmic bias and policing see Marion Oswald & Alexander Babuta, *Data Analytics and Algorithmic Bias in Policing* (2019); Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153–163 (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 165 (2011); on this issue in India see Vidushi Marda & Shivangi Narayan, *Data in New Delhi's predictive policing system*, in PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 317–324 (2020), <https://doi.org/10.1145/3351095.3372865> (last visited Jan 9, 2021).

¹³ For more discussion on bias and EdTech see Elana Zeide, *The Structural Consequences of Big Data-Driven Education*, 5 BIG DATA 164–172 (2017); Elana Zeide, *The Limits of Education Purpose Limitations*, 71 U. MIAMI L. REV. 494 (2016); Priscilla M. Regan & Jolene Jesse, *Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking*, 21 ETHICS INF TECHNOL 167–179 (2019).

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 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. 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. 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.¹⁴ 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 racial¹⁵ and gender¹⁶ 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

¹⁴ 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 recognise that social barriers and inequalities manifest differently across different countries, and that specific barriers found in certain countries (e.g. the USA) 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 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.

¹⁵ JEAN HALLEY, AMY ESHLEMAN & RAMYA MAHADEVAN VIJAYA, *SEEING WHITE: AN INTRODUCTION TO WHITE PRIVILEGE AND RACE* 120–121, 127, 136 (2011).

¹⁶ ANGELA SAINI, *INFERIOR: HOW SCIENCE GOT WOMEN WRONG AND THE NEW RESEARCH THAT’S REWRITING THE STORY* 9–11 (2017).

intelligence tests, have been shown to be racially biased against minority groups including Roma, Black people, Latinx, and migrants.¹⁷ Encouragement and assessments of Jade's academic merit by her teachers and professors may also have been shaded by racial¹⁸ and gender bias,¹⁹ and ultimately reflected in her marks or reference letters from educators.²⁰ Similar gender biases may have influenced her male peers;²¹ children have been shown, for example, to already develop clear ideas of gender roles by the age of six.²²

¹⁷ With regards to Roma and the EU, see: Case 57325/00 ECHR, D.H. and Others v. the Czech Republic, 2007, <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. Case 57325/00 ECHR, D.H. and Others v. the Czech Republic, 2007, <https://hudoc.echr.coe.int/eng?i=001-83256>; For more details see I. Chopin, C. Germaine & J. Tanczos, *Roma and the enforcement of anti-discrimination law*, LUXEMBOURG: EUROPEAN NETWORK OF LEGAL EXPERTS IN GENDER EQUALITY AND NON-DISCRIMINATION, 13–18 (2017); CLAUDE S. FISCHER ET AL., *INEQUALITY BY DESIGN: CRACKING THE BELL CURVE MYTH*, 172–173 (1996); With regards to the US and Black and Latinx people and immigrants, see: HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 40.

¹⁸ See a study in the UK Simon Burgess & Ellen Greaves, *Test scores, subjective assessment, and stereotyping of ethnic minorities*, 31 JOURNAL OF LABOR ECONOMICS 535–576 (2013). RENI EDDO-LODGE, *WHY I'M NO LONGER TALKING TO WHITE PEOPLE ABOUT RACE* 66–67 (2020). With regards to mathematics skill versus received marks and encouragement based on ethnicity, see: Rickie Sanders, *Gender equity in the classroom: An arena for correspondence*, 28 WOMEN'S STUDIES QUARTERLY 182–193 (2000).

¹⁹ As above concerning mathematics skill versus grades. The grades of the boys 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. See VICTOR LAVY & EDITH SAND, *On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases* (2015); Sue Wilson, *Teachers' gender bias in maths affects girls later*, THE CONVERSATION, <http://theconversation.com/teachers-gender-bias-in-maths-affects-girls-later-37844> (last visited Sep 20, 2020).

²⁰ 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 KuoOct. 3, 2016 & 12:00 Pm, *Recommendation letters reflect gender bias*, SCIENCE | AAAS (2016), <https://www.sciencemag.org/careers/2016/10/recommendation-letters-reflect-gender-bias> (last visited Sep 20, 2020); CAROLINE CRIADO PEREZ, *INVISIBLE WOMEN: EXPOSING DATA BIAS IN A WORLD DESIGNED FOR MEN* 102 (2019). See also ADVANCE-HE, *Equality in higher education: statistical report 2013 Part 2: students* 80, https://s3.eu-west-2.amazonaws.com/assets.creode.advancehe-document-manager/documents/ecu/equality-in-he-statistical-report-2013-students_1579016961.pdf (last visited Dec 17, 2020).

²¹ Daniel Z. Grunspan et al., *Males Under-Estimate Academic Performance of Their Female Peers in Undergraduate Biology Classrooms*, 11 PLOS ONE e0148405 (2016).

²² 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. 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. See Carol Lynn Martin & Diane Ruble, *Children's search for gender cues: Cognitive perspectives on gender development*, 13 CURRENT DIRECTIONS IN PSYCHOLOGICAL SCIENCE 67–70, 68 (2004).

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 outcomes²³ and increase the inequality experienced by historically disadvantaged groups.²⁴

Recognising this problem, much work has emerged in recent years to address bias in machine learning and AI systems.²⁵ Many scholars urge for greater accountability in their design and usage.²⁶

²³ For seminal work on the widening of existing inequality see CATHY O’NEIL, *WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY* (2017); and VIRGINIA EUBANKS, *AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR* (2018).

²⁴ O’NEIL, *supra* note 23; EUBANKS, *supra* note 23.

²⁵ O’NEIL, *supra* note 23; Brent Mittelstadt et al., *The ethics of algorithms: Mapping the debate*, 3 *BIG DATA & SOCIETY* (2016), <http://bds.sagepub.com/lookup/doi/10.1177/2053951716679679> (last visited Dec 15, 2016); EUBANKS, *supra* note 23; Tal Zarsky, *The Trouble with Algorithmic Decisions An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 *SCIENCE TECHNOLOGY HUMAN VALUES* 118–132 (2016); Solon Barocas et al., *The problem with bias: Allocative versus representational harms in machine learning*, in 9TH ANNUAL CONFERENCE OF THE SPECIAL INTEREST GROUP FOR COMPUTING, INFORMATION AND SOCIETY (2017); Jenna Burrell, *How the Machine “Thinks:” Understanding Opacity in Machine Learning Algorithms*, *BIG DATA & SOCIETY* (2016); Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics derived automatically from language corpora contain human-like biases*, 356 *SCIENCE* 183–186 (2017); Timnit Gebru et al., *Datasheets for datasets*, ARXIV PREPRINT ARXIV:1803.09010 (2018); Margaret Mitchell et al., *Model cards for model reporting*, in PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 220–229 (2019); DEVEN R. DESAI & JOSHUA A. KROLL, *Trust But Verify: A Guide to Algorithms and the Law* (2017), <https://papers.ssrn.com/abstract=2959472> (last visited Jan 29, 2021).

²⁶ To name only a few VIKTOR MAYER-SCHÖNBERGER & THOMAS RAMGE, *REINVENTING CAPITALISM IN THE AGE OF BIG DATA* (2018); Ari Ezra Waldman, *Power, Process, and Automated Decision-Making*, 88 *FORDHAM L. REV.* 613 (2019); KAREN YEUNG & MARTIN LODGE, *ALGORITHMIC REGULATION* (2019); Omer Tene & Jules Polonetsky, *Taming the Golem: Challenges of ethical algorithmic decision-making*, 19 *NCJL & TECH.* 125 (2017); Jatinder Singh et al., *Responsibility & machine learning: Part of a process*, AVAILABLE AT SSRN 2860048 (2016); David Lehr & Paul Ohm, *Playing with the data: what legal scholars should learn about machine learning*, 51 *UCDL REV.* 653 (2017); Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 *COLUMBIA LAW REVIEW* 1851–1886 (2019); B. Bodo 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); Jatinder Singh, Jennifer Cobbe & Chris Norval, *Decision Provenance: Harnessing data flow for accountable systems*, 7 *IEEE ACCESS* 6562–6574 (2019); Christian Sandvig et al., *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> (last visited Feb 13, 2016); ANDREW TUTT, *An FDA for Algorithms* (2016), <http://papers.ssrn.com/abstract=2747994> (last visited Apr 13, 2016); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 *UCLA L. REV.* 54 (2019); Sara Hajian, Francesco Bonchi & Carlos Castillo, *Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining*, in PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 2125–2126 (2016), <https://doi.org/10.1145/2939672.2945386> (last visited Jan 9, 2021); Indrė Žliobaitė, *Measuring discrimination in algorithmic decision making*, 31 *DATA MINING AND KNOWLEDGE DISCOVERY* 1060–1089 (2017);

Machine learning systems take in data and produce outputs, such as decisions or classifications, based on learned rules or parameters.²⁷ 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.²⁸

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’.²⁹ 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.³⁰ 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.

Comparatively speaking, ‘social biases’ are very difficult to ‘fix’.³¹ 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.³² 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

Laurens Naudts, *How Machine Learning Generates Unfair Inequalities and How Data Protection Instruments May Help in Mitigating Them*, R. LEENES, R. VAN BRAKEL, S. GUTWIRTH & P. DE HERT (AUTHORS), DATA PROTECTION AND PRIVACY: THE INTERNET OF BODIES (COMPUTERS, PRIVACY AND DATA PROTECTION (2019).

²⁷ See S. C. Olhede & P. J. Wolfe, *The growing ubiquity of algorithms in society: implications, impacts and innovations*, 376 PHIL. TRANS. R. SOC. A 20170364 (2018); JOSHUA A. KROLL ET AL., *Accountable Algorithms* (2016), <http://papers.ssrn.com/abstract=2765268> (last visited Apr 29, 2016).

²⁸ This simplified view is inspired by Batya Friedman & Helen Nissenbaum, *Bias in computer systems*, 14 ACM TRANSACTIONS ON INFORMATION SYSTEMS (TOIS) 330–347, 333–335 (1996); see also Mireille Hildebrandt, *The Issue of Bias. The Framing Powers of ML* (2020).

²⁹ 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 and Nissenbaum, *supra* note 28 at 333–335.

³⁰ *Id.* at 333–335.

³¹ 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: Section 3). This definition roughly follows the taxonomy proposed by *Id.* 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.’ However, breaking with their taxonomy, we attribute biases arising from both individual and societal sources as “social bias.”

³² *Id.* at 334.

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.³³

Recognising 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.³⁴ Non-intervention frequently amplifies and widens existing inequalities in our society that have been learned by a model 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,³⁵ 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.³⁶ This is a

³³ Among others see Timnit Gebru, *Oxford Handbook on AI Ethics Book Chapter on Race and Gender*, ARXIV PREPRINT ARXIV:1908.06165 (2019); Ruha Benjamin, *Race after technology: Abolitionist tools for the new jim code*, SOCIAL FORCES (2019); CATHERINE D’IGNAZIO & LAUREN F. KLEIN, DATA FEMINISM (2020); Rediet Abebe et al., *Roles for computing in social change*, in PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 252–260 (2020); Chelsea Barabas et al., *Interventions over predictions: Reframing the ethical debate for actuarial risk assessment*, in CONFERENCE ON FAIRNESS, ACCOUNTABILITY AND TRANSPARENCY 62–76 (2018); Cynthia L. Bennett & Os Keyes, *What is the point of fairness? disability, ai and the complexity of justice*, in ASSETS 2019 WORKSHOP—AI FAIRNESS FOR PEOPLE WITH DISABILITIES (2019); Daniel Greene, Anna Lauren Hoffmann & Luke Stark, *Better, nicer, clearer, fairer: 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 et al., *Questioning the assumptions behind fairness solutions*, ARXIV PREPRINT ARXIV:1811.11293 (2018); Shira Mitchell et al., *Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions*, ARXIV PREPRINT ARXIV:1811.07867 (2018).

³⁴ See O’NEIL, *supra* note 23; and EUBANKS, *supra* note 23.

³⁵ 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 also Anna Lauren Hoffmann, *Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse*, 22 INFORMATION, COMMUNICATION & SOCIETY 900–915, 907 (2019); in favour, see: Abigail Z. Jacobs & Hanna Wallach, *Measurement and Fairness*, ARXIV:1912.05511 [CS] (2021), <http://arxiv.org/abs/1912.05511> (last visited Jan 21, 2021).

³⁶ 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

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 to not 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’³⁷ fairness metrics) aligns with a different fundamental normative concept in EU non-discrimination law: substantive equality. According to substantive equality or “*de facto* equality”³⁸, 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 paper 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 (‘bias preserving’ and ‘bias transforming’ fairness metrics) and how well they align with the aims of non-discrimination law. Finally, we recognise that the legal need for justification in cases of indirect discrimination may create new obligations for developers, deployers, and users. Recognising this need for justification, we argue that metrics which require an explicit choice to be made of which biases a classifier should inherit should be preferred for purposes of fair decision-making under substantive equality. To conclude, we give concrete

tool that can be used for better or worse, picking a seemingly neutral baseline such as maintaining the status quo seems intuitively sensible.

³⁷ 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. As a result, 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* (2020), <https://papers.ssrn.com/abstract=3547922> (last visited Apr 19, 2020).

³⁸ Marianne Henriette Simone Gijzen, *Selected issues in equal treatment law: a multi-layered comparison of European, English and Dutch law*, 23 (2006).

recommendations how to choose the most appropriate fairness metric under EU non-discrimination law and a checklist to do so.

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 (and 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 mean that all legal issues are resolved. 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 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 note of context is necessary. Our intention in this paper 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 bias, and (2) as a basis for making fair decisions in practice. Testing for and mitigating technical bias remains a vital area of research.³⁹ Similarly, when making decisions about individuals in cases where explicit normative decisions have not yet been taken regarding which bias the system should exhibit, technical bias mitigation helps ensure that systems are not making things worse. In scenarios where either ‘ground truth’ labels can be exactly known⁴⁰, no bias exists,⁴¹ known disparity has been previously legally justified, or

³⁹ 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.

⁴⁰ For example, predicting the outcome of a preassigned biopsy test.

⁴¹ 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

where systems need to be designed to replicate societal bias (e.g. as a diagnostical 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, bias preserving metrics might also be sufficient 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 additional 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 proactively provide an objective justification for the use of bias preserving metrics.

2 FORMAL AND SUBSTANTIVE EQUALITY IN NON-DISCRIMINATION LAW

EU non-discrimination law prohibits two types of discrimination: direct and indirect discrimination.⁴² Direct discrimination means that a person is treated less favourably based on a protected attribute (e.g. race and ethnicity,⁴³ gender,⁴⁴ religion and belief, age, disability, or sexual orientation⁴⁵) that they possess in matters of a

repay?” To gather this data without sampling bias a randomized control trial would be needed. See also Mitchell et al., *supra* note 33; Bernhard Schölkopf, *Causality for machine learning*, ARXIV PREPRINT ARXIV:1911.10500 (2019).

⁴² In general see MARK BELL, *ANTI-DISCRIMINATION LAW AND THE EUROPEAN UNION* (2002); for an assessment of scope limitation see Justyna Maliszewska-Nienartowicz, *Direct and indirect discrimination in European union law—how to draw a dividing line*, 3 INTERNATIONAL JOURNAL OF SOCIAL SCIENCES 41–55 (2014); Erica Howard, *EU anti-discrimination law: Has the CJEU stopped moving forward?*, 18 INTERNATIONAL JOURNAL OF DISCRIMINATION AND THE LAW 60–81 (2018).

⁴³ EUROPEAN COUNCIL, *Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin* (2000), <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32000L0043>.

⁴⁴ DIRECTIVE 2006/54/EC 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), , OJ L 204 (2006), <http://data.europa.eu/eli/dir/2006/54/oj/eng> (last visited Aug 5, 2019); EUROPEAN PARLIAMENT’S DIRECTORATE-GENERAL FOR PARLIAMENTARY RESEARCH SERVICES, *Gender Equal Access to Goods and Services Directive 2004/113/EC -European Implementation Assessment*, [http://www.europarl.europa.eu/RegData/etudes/STUD/2017/593787/EPRS_STU\(2017\)593787_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/STUD/2017/593787/EPRS_STU(2017)593787_EN.pdf) (last visited Mar 26, 2019).

⁴⁵ COUNCIL DIRECTIVE 2000/78/EC OF 27 NOVEMBER 2000 ESTABLISHING A GENERAL FRAMEWORK FOR EQUAL TREATMENT IN EMPLOYMENT AND OCCUPATION, , OJ L 303 (2000), <http://data.europa.eu/eli/dir/2000/78/oj/eng> (last visited Aug 5, 2019); 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* 13–36 (2017); *GROUP PRIVACY: NEW CHALLENGES OF DATA*

protected sector (e.g. the workplace, provision of goods and services).⁴⁶ Different groups receive different levels of protection.⁴⁷ 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.”⁴⁸

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

TECHNOLOGIES, (Linnet Taylor, Luciano Floridi, & Bart van der Sloot eds., 1 ed. 2017); Brent Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, 30 PHILOSOPHY & TECHNOLOGY 475–494 (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 139–158 (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 TECHNOLOGY LAW JOURNAL (2020), <https://papers.ssrn.com/abstract=3388639> (last visited Feb 9, 2020); on how AI fairness methodologies fail to adequately account for the socially constructed nature of groups such as race see Alex Hanna et al., *Towards a critical race methodology in algorithmic fairness*, in PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 501–512 (2020), <https://doi.org/10.1145/3351095.3372826> (last visited Jan 8, 2021).

⁴⁶ For details on scope and limitations see WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37; Philipp Hacker, *Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law*, 55 COMMON MARKET LAW REVIEW 1143–1185 (2018); and Frederik J. Zuiderveen Borgesius, *Strengthening legal protection against discrimination by algorithms and artificial intelligence*, THE INTERNATIONAL JOURNAL OF HUMAN RIGHTS 1–22 (2020); Dimitri Droshout et al., *Non-discrimination law*, HART: PORTLAND (2007); DAGMAR SCHIEK ET AL., CASES, MATERIALS AND TEXT ON NATIONAL, SUPRANATIONAL AND INTERNATIONAL NON-DISCRIMINATION LAW: IUS COMMUNE CASEBOOKS FOR THE COMMON LAW OF EUROPE (2007); for a comparative view see Raphaël Gellert et al., *A comparative analysis of anti-discrimination and data protection legislations*, in DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 61–89 (2013); on the scope of the European Convention of Human Rights see F. Zuiderveen Borgesius, *Discrimination, artificial intelligence, and algorithmic decision-making* (2018); Janneke Gerards, *The discrimination grounds of Article 14 of the European Convention on Human Rights*, 13 HUMAN RIGHTS LAW REVIEW 99–124 (2013); Mark Bell, *The Implementation of European Anti-Discrimination Directives: Converging towards a Common Model?*, 79 THE POLITICAL QUARTERLY 36–44 (2008); *Id.*

⁴⁷ For a discussion on group hierarchy see Erica Howard, *The Case for a Considered Hierarchy of Discrimination Grounds in EU Law*, 13 MAASTRICHT JOURNAL OF EUROPEAN AND COMPARATIVE LAW 445–470 (2006).

⁴⁸ 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 EUROPEAN COMMISSION, DIRECTORATE-GENERAL FOR EMPLOYMENT, SOCIAL AFFAIRS AND EQUAL OPPORTUNITIES, UNIT G (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*, HARVARD LAW REVIEW 537–596 (1982).

maintain and exacerbate inequalities and inhibit substantive equality of opportunity in practice.⁴⁹

In the words of President Lyndon 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,” 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.”⁵⁰

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”⁵¹), cannot be achieved simply by ignoring protected attributes (e.g. race, gender, disability) and treating everyone the same going forward. A more active attitude is required that accounts for social and historical realities. 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.⁵²

A useful distinction can be drawn between procedural and substantive equal opportunity.⁵³ 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

⁴⁹ Catharine A. MacKinnon, *Toward a renewed equal rights amendment: Now more than ever*, 37 HARV. JL & 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 AND WATSON, *supra* note 48.

⁵⁰ Lyndon Johnson, Howard University Commencement Address (1965) | The American Yawp Reader, , <https://www.americanyawp.com/reader/27-the-sixties/lyndon-johnson-howard-university-commencement-address-1965/> (last visited Nov 7, 2020); For an overview of the history, aims and limitations of UK and US non-discrimination law see Christopher McCrudden, *Institutional discrimination*, 2 OXFORD J. LEGAL STUD. 303 (1982).

⁵¹ Gijzen, *supra* note 38 at 23.

⁵² STIGLITZ, *supra* note 11 at 199; Hoffmann, *supra* note 35 at 904; EDDO-LODGE, *supra* note 18 at 63.

⁵³ Bernard Williams, *The Idea of Equality*, *Philosophy*, 2 POLITICS, AND SOCIETY, SERIES, 125–126 (1962); in favor of Williams see Sandra Fredman, *Substantive equality revisited*, 14 INTERNATIONAL JOURNAL OF CONSTITUTIONAL LAW 712–738, 723–724 and 735 (2016); SANDRA FREDMAN, *DISCRIMINATION LAW* (2011).

existing patterns of disadvantage.⁵⁴ Here the focus is more on restructuring society rather than the individual, even though formal equal opportunity as an interim step has tremendous value.⁵⁵

2.1 Indirect discrimination and substantive equality

The concept of indirect discrimination was created to achieve substantive equality in practice.⁵⁶ 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 [...]”⁵⁷ Indirect discrimination is intended to help redistribute resources from the advantaged to the disadvantaged and to promote diversity in society.⁵⁸ It enables non-discrimination law to play a more active role in creating substantive equality by tackling subtle social and historical inequalities.⁵⁹

Indirect discrimination occurs when a “apparently neutral provision, criterion or practice”⁶⁰ 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 (at a minimum) who are, on average, shorter than men.⁶¹ Elsewhere, the authors have argued that indirect discrimination is the most likely type of discrimination to

⁵⁴ There are a multitude of theories of equality of opportunity. For discussion of its conception within EU non-discrimination law, see Williams, *supra* note 53 at 125–126; Fredman, *supra* note 53 at 723–724 and 735; FREDMAN, *supra* note 53; For discussion and criticism of its varied manifestations in philosophy see for example Gerald A. Cohen, *On the currency of egalitarian justice*, 99 ETHICS 906–944 (1989); JOHN E. ROEMER, EQUALITY OF OPPORTUNITY (2009); Elizabeth S. Anderson, *What Is the Point of Equality?*, 109 ETHICS 287–337 (1999); ANDREW MASON, LEVELLING THE PLAYING FIELD: THE IDEA OF EQUAL OPPORTUNITY AND ITS PLACE IN EGALITARIAN THOUGHT (2006), <https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199264414.001.0001/acprof-9780199264414> (last visited Feb 25, 2021).

⁵⁵ EDDO-LODGE, *supra* note 18 at 184.

⁵⁶ LIMITS AND POTENTIAL OF THE CONCEPT OF INDIRECT DISCRIMINATION, 24 (Christa Tobler & Europäische Kommission eds., Ms. completed in September 2008 ed. 2008).

⁵⁷ *Id.* at 24. citing; Gijzen, *supra* note 38 at 82; Dagmar Schiek, *Indirect discrimination*, CASES, MATERIALS AND TEXT ON NATIONAL, SUPRANATIONAL AND INTERNATIONAL NON-DISCRIMINATION LAW. OXFORD: HART PUBLISHING 323–476, 327 (2007).

⁵⁸ Gijzen, *supra* note 38 at 136.

⁵⁹ Marc De Vos, *The European Court of Justice and the march towards substantive equality in European Union anti-discrimination law*, 20 INTERNATIONAL JOURNAL OF DISCRIMINATION AND THE LAW 62–87, 72 (2020); Sandra Fredman, *Equality: a new generation?*, 30 INDUSTRIAL LAW JOURNAL 145–168, 161 (2001); LIMITS AND POTENTIAL OF THE CONCEPT OF INDIRECT DISCRIMINATION, *supra* note 56 at 24.

⁶⁰ This 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?*, EUROPEAN EQUALITY LAW REVIEW, FORTHCOMING, at 3 (2016).

⁶¹ This example is inspired by German case around height requirements for pilots. See WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 12.

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.⁶²

Once a claimant establishes *prima facie* indirect discrimination in court, the burden of proof shifts to the alleged offender.⁶³ The alleged offender then has 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.⁶⁴

Indirect discrimination differs from direct discrimination by acknowledging that the social hurdles, struggles, and factual differences facing protected groups must be taken into consideration.⁶⁵ 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 favourable changes in part-time work and employment law, which can be justified under the proportionality test.⁶⁶

2.1.1 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.⁶⁷ Positive action, whilst also a form of substantive equality,⁶⁸ focuses solely on equality of

⁶² WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37; Wachter and Mittelstadt, *supra* note 7.

⁶³ More on this see Julie Ringelheim, *The Burden of Proof in Antidiscrimination Proceedings. A Focus on Belgium, France and Ireland*, A FOCUS ON BELGIUM, FRANCE AND IRELAND (SEPTEMBER 4, 2019), EUROPEAN EQUALITY LAW REVIEW (2019); on the practical limitations see LILLA FARKAS ET AL., *Reversing the burden of proof: practical dilemmas at the European and national level* (2015), <http://dx.publications.europa.eu/10.2838/05358> (last visited Feb 9, 2020).

⁶⁴ For an extensive overview of EU case law on reasons to justify discrimination, procedural norms, and admissible evidence see Wachter, *supra* note 45 at 46–54; WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37.

⁶⁵ De Vos, *supra* note 59 at 72.

⁶⁶ *Id.* at 74.; Sandra Fredman, *Addressing disparate impact: Indirect discrimination and the public sector equality duty*, 43 INDUSTRIAL LAW JOURNAL 349–363, 363 (2014).

⁶⁷ ELLIS AND WATSON, *supra* note 48 at 176–177 and Chapter 9.

⁶⁸ 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 US context see, see: 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 53 at 713; Thomas Nachbar, *Algorithmic Fairness, Algorithmic Discrimination*, VIRGINIA PUBLIC LAW AND LEGAL THEORY RESEARCH PAPER, 34 (2020); Solon Barocas & Andrew D. Selbst, *Big data’s disparate impact*, 104 CALIFORNIA LAW REVIEW, 726 (2016); Pauline T Kim, *Data-Driven Discrimination*

outcomes.⁶⁹ Adjusting the outcomes of a 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.⁷⁰ 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. not relying heavily on recommendation letters or GPA). Substantive equality is satisfied when everybody starts ‘the race’ from the same point (e.g. via equal access to education or healthcare), and not only when a specific number of people from a certain group have won the race.⁷¹ Indirect discrimination “diagnoses discrimination,”⁷² 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.⁷³

2.2 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.⁷⁴ As Ellis and Watson argue, “[i]f the moral basis on which the law forbids

at Work, 58 81; James Grimmelman & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. CIRCUIT 164 (2016); Zachary Lipton, Julian McAuley & Alexandra Chouldechova, *Does mitigating ML’s impact disparity require treatment disparity?*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 8125–8135 (2018); Crystal Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, AVAILABLE AT SSRN 3462379 (2019); Michael Feldman et al., *Certifying and Removing Disparate Impact*, in PROCEEDINGS OF THE 21TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 259–268 (2015), <https://doi.org/10.1145/2783258.2783311> (last visited Jan 8, 2021); CATHARINE A. MACKINNON, BUTTERFLY POLITICS (2017); for seminal work on US non-discrimination law see MARTHA MINOW, MAKING ALL THE DIFFERENCE: INCLUSION, EXCLUSION, AND AMERICAN LAW (1990).

⁶⁹ 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).

⁷⁰ ELLIS AND WATSON, *supra* note 48 at 7.

⁷¹ Fredman, *supra* note 53 at 723 and 729.

⁷² Fredman, *supra* note 59 at 161.

⁷³ 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*, AVAILABLE AT SSRN 3524238, 220 (2019).

⁷⁴ 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 53 at 713.

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[...].”⁷⁵ 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.⁷⁶ 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;⁷⁷ 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.”⁷⁸

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.⁷⁹ 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.⁸⁰

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.⁸¹ In cases which meet the strict

⁷⁵ ELLIS AND WATSON, *supra* note 48 at 4.

⁷⁶ Fredman, *supra* note 59 at 156 also refers to human dignity as a principle of equality law to prevent levelling down to achieve parity (i.e. treating everyone equally bad).

⁷⁷ AMARTYA SEN, DEVELOPMENT AS FREEDOM (2001); MARTHA NUSSBAUM, WOMEN AND HUMAN DEVELOPMENT: A STUDY IN HUMAN CAPABILITIES (2000).

⁷⁸ NUSSBAUM, *supra* note 77 at 90–91.

⁷⁹ See evolving case law 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-555/07, Seda Küçükdeveci v Swedex GmbH & Co. KG., 2010 E.C.R. I-21, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62007CJ0555> (last visited Aug 7, 2019); Case 109/88, Handels- og Kontorfunktionærernes Forbund I Danmark v Dansk Arbejdsgiverforening, acting on behalf of Danfoss, 1989 E.C.R. I-03199, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A61988CJ0109>; 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>; Joined Cases C-569/16 and C-570/16, Stadt Wuppertal v Maria Elisabeth Bauer and Volker Willmeroth v Martina Broßonn, 2018 E.C.R. I-871, <http://curia.europa.eu/juris/liste.jsf?num=C-569/16&language=en> (last visited Jan 13, 2021).

⁸⁰ Article 21 states that “[a]ny discrimination based on any ground such as sex, race, colour, 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).

⁸¹ See De Vos, *supra* note 59 at 81; Among other cases, DeVos cites the following cases as evidence that the ECJ sees substantive equality as the aim of the law Case C-54/07, Centrum

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.

2.3 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,⁸² Fraser and Honneth,⁸³ Collins,⁸⁴ and Barnard⁸⁵ the goal 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 legislator and the Court of Justice for the lack of clear definition of those substantive equality goals and aims.⁸⁶

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?⁸⁷

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 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, *supra* note 79; 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, *supra* note 79; 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 LJ 562 (2000).

⁸² Fredman, *supra* note 53 at 732.

⁸³ NANCY FRASER & AXEL HONNETH, REDISTRIBUTION OR RECOGNITION?: A POLITICAL-PHILOSOPHICAL EXCHANGE 36–37 (2003).

⁸⁴ Hugh Collins, *Discrimination, equality and social inclusion*, 66 THE MODERN LAW REVIEW 16–43, 24 (2003).

⁸⁵ Catherine Barnard, *The future of equality law: equality and beyond*, THE FUTURE OF LABOUR LAW. LIBER AMICORUM FOR BOB HEPPLER, OXFORD: HART (2004).

⁸⁶ De Vos, *supra* note 59 at 83.

⁸⁷ For discourse on how best to achieve substantive equality see MacKinnon, *supra* note 68; Fredman, *supra* note 53; Catharine A. MacKinnon, *Substantive equality revisited: A reply to Sandra Fredman*, 14 INTERNATIONAL JOURNAL OF

To achieve equality of distribution of goods for all? To accommodate diversity?⁸⁸

- What role (passive or active) is expected of the regulator, the legislator and the private and public sector?⁸⁹
- Should there be a practice and pre-emptive duty of the public and the private sector to dismantle inequality?⁹⁰
- Can this happen at the expense of dominant groups, potentially leading to positive discrimination?⁹¹
- When can disparity be legally justified?⁹²
- How should the law address intersectional discrimination?⁹³

The extent to which the law does and should impose positive obligations on public and private actors to achieve substantive equality is a particularly difficult question.⁹⁴ Positive obligations could require actors to, for example, actively promote equal opportunity or even redistribute resources or jobs.⁹⁵ Negative obligations could restrict an actor's ability to base decisions on criteria that are known to disproportionately disadvantage certain protected groups.⁹⁶

CONSTITUTIONAL LAW 739–746 (2016); Sandra Fredman, *Substantive equality revisited: A rejoinder to Catharine MacKinnon*, 14 INTERNATIONAL JOURNAL OF CONSTITUTIONAL LAW 747–751 (2016); Catharine A MacKinnon, *Substantive equality revisited: A rejoinder to Sandra Fredman*, 15 INTERNATIONAL JOURNAL OF CONSTITUTIONAL LAW 1174–1177 (2017) The dialogue discusses whether framing equality around traditional power hierarchies or around multifaceted ways of oppression is better to support substantive equality.

⁸⁸ On this see also Fredman, *supra* note 59 at 164–165.

⁸⁹ Urs Gasser & Carolyn Schmitt, *The Role of Professional Norms in the Governance of Artificial Intelligence* (2019); Technology | Academics | Policy - Jonathan Zittrain and Jack Balkin Propose Information Fiduciaries to Protect Individual Privacy Rights, <http://www.techpolicy.com/Blog/September-2018/Jonathan-Zittrain-and-Jack-Balkin-Propose-Informat.aspx> (last visited Feb 2, 2019).

⁹⁰ In favour see SANDRA FREDMAN, EUROPEAN COMMISSION & EUROPEAN NETWORK OF LEGAL EXPERTS IN THE FIELD OF GENDER EQUALITY, *Making Equality Effective: The role of proactive measures* (2009).

⁹¹ For an account clearly in favour of whether this measure meets the EU's proportionality test see Fredman, *supra* note 66 at 363.

⁹² Wachter, *supra* note 45 at 46–54.

⁹³ Kimberle Crenshaw, *Mapping the margins: Intersectionality, identity politics, and violence against women of color*, 43 STAN. L. REV. 1241 (1990); Devon W. Carbado et al., *Intersectionality: Mapping the movements of a theory*, 10 DU BOIS REVIEW: SOCIAL SCIENCE RESEARCH ON RACE 303–312 (2013); on the lack of legal protection under EU law see WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 20.

⁹⁴ For an overview of measures in the EU Member States, see FREDMAN, EUROPEAN COMMISSION, AND EUROPEAN NETWORK OF LEGAL EXPERTS IN THE FIELD OF GENDER EQUALITY, *supra* note 90.

⁹⁵ ELLIS AND WATSON, *supra* note 48 at 7.

⁹⁶ 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 marginalised groups access to financial services? Similarly, under what circumstances can disparity be justified? For example, should a bank be able

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.⁹⁷ Nonetheless, several possible grounds for broad positive duties have been identified.

In the UK, for example, the justification defence 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.⁹⁸ 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.⁹⁹ However, a clear and general legal duty for preventative, positive duty is not yet established. Nonetheless, De Vos,¹⁰⁰ Tobler,¹⁰¹ and Fredman¹⁰² 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.¹⁰³ 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

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.

⁹⁷ De Vos, *supra* note 59 at 72–73. Fredman, *supra* note 53 at 723–724 and 735; FREDMAN, *supra* note 53. For example, cases dealing with discrimination based on disability recognise a duty for “reasonable accommodation.”

⁹⁸ Fredman, *supra* note 66 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 defence also exists.

⁹⁹ 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). Complaints often do not deter perpetrators. Proactive measures on the other hand would benefit everyone and not just the claimant and would contribute to the systematic erosion of inequality. FREDMAN, EUROPEAN COMMISSION, AND EUROPEAN NETWORK OF LEGAL EXPERTS IN THE FIELD OF GENDER EQUALITY, *supra* note 90 at 1–5 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).

¹⁰⁰ “A limited duty of preventive positive action is therefore implicit in the prohibition of indirect discrimination.” See De Vos, *supra* note 59 at 71; see also MARK DE VOS, BEYOND FORMAL EQUALITY: POSITIVE ACTION UNDER DIRECTIVES 2000/43/EC AND 2000/78/EC 81 (2007).

¹⁰¹ LIMITS AND POTENTIAL OF THE CONCEPT OF INDIRECT DISCRIMINATION, *supra* note 56 at 92.

¹⁰² Fredman, *supra* note 59 at 164 and see page 167 “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.”; Fredman, *supra* note 53 at 735.

¹⁰³ Fredman, *supra* note 59 at 164; Fredman, *supra* note 53 at 735.

occurred the outcome should not just be individual compensation, but a requirement of restructuring.¹⁰⁴ 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.¹⁰⁵ Other legal remedies to bolster the complaint-based system and lessen the burden of the individuals include the strengthening of oversight power of equality bodies¹⁰⁶ and collective redress mechanisms.¹⁰⁷

When dealing with scarce resources this restructuring can of course mean that traditional benefactors lose out. However, EU law¹⁰⁸ and case law¹⁰⁹ support measures of equal opportunity as well as positive action (with some restrictions)¹¹⁰ to support societal and systemic restructuring. This holds true even if restructuring means that historically dominant groups receive less favourable treatment and (potential) positive discrimination occurs.¹¹¹ Both can be justified under the EU's proportionality test.¹¹²

3 BIAS PRESERVATION IN FAIR MACHINE LEARNING

The existence and precise requirements imposed by positive substantive equality duties for the public and private sectors 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

¹⁰⁴ Fredman, *supra* note 59 at 163.

¹⁰⁵ FREDMAN, EUROPEAN COMMISSION, AND EUROPEAN NETWORK OF LEGAL EXPERTS IN THE FIELD OF GENDER EQUALITY, *supra* note 90 at 5.

¹⁰⁶ 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).

¹⁰⁷ Sara Benedi Lahuerta, *Enforcing EU equality law through collective redress: Lagging behind?*, 55 COMMON MARKET LAW REVIEW (2018).

¹⁰⁸ Fredman, *supra* note 66 at 363.

¹⁰⁹ De Vos, *supra* note 59 at 74–75.

¹¹⁰ For case law on this issue see McCrudden, *supra* note 73 at 220.

¹¹¹ “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.” See De Vos, *supra* note 59 at 81.

¹¹² Fredman, *supra* note 66 at 363.

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.¹¹³ The bias could be rejected for bringing about negative consequences, or violating some fundamental ethical principle, or simply contravening legal provisions against direct discrimination (see: Section 2).

As is typically the case in fair machine learning research, throughout this paper 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.¹¹⁴ 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.¹¹⁵

3.1 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.¹¹⁶ 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.¹¹⁷

Formally, the definition of demographic parity asserts that each protected group, meaning a group based on a protected attribute

¹¹³ 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.

¹¹⁴ More precisely, in this paper 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.

¹¹⁵ Throughout this paper we assume that the reader finds at least some of the inequality that currently exists in the world normatively problematic.

¹¹⁶ WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37.

¹¹⁷ *Id.* at 48–54. We also emphasise 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.

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, the authors have previously examined justification of apparent indirect discrimination,¹¹⁸ by which a practice that would otherwise be considered discriminatory can be justified on the basis of a legitimate interest and the proportionality test (see: Section 2.1).¹¹⁹ 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 defence 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).¹²⁰

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 recognised as an explicit dependence on target labels (see: Section 3.2). 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?

¹¹⁸ *Id.* at 41–44.; Wachter, *supra* note 45 at 46–54.

¹¹⁹ For more detail on the ECJ case law on what can be objectively justified see Wachter, *supra* note 45 at 41–44. Further, examples can be seen in the application of the UK Equality Act 2010 that transposes the EU Non-Discrimination Directives. Here the UK Equality and Human Rights Commission offers the following guidance on requirements for an objective justification that would satisfy the courts. 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 unfavourable 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.” See Equality and Human Rights Commission, *Words and terms used in the Equality Act*, <https://www.equalityhumanrights.com/en/advice-and-guidance/commonly-used-terms-equal-rights> (last visited Dec 10, 2020).

¹²⁰ WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37.

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.¹²¹ 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.

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.¹²² 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.

3.2 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

¹²¹ Moritz Hardt, Eric Price & Nati Srebro, *Equality of opportunity in supervised learning*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 3315–3323 (2016).

¹²² WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 13–32.

not necessarily satisfied by a perfect classifier, we refer to as *bias transforming*.¹²³

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 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,¹²⁴ equal opportunity,¹²⁵ and calibration,¹²⁶ implicitly assume that various forms of bias in the historic data Y are there for a reason and should be preserved.¹²⁷

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.¹²⁸

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 that requires justification when prima facie discrimination is established,¹²⁹ 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 (see: Section 5).

¹²³ As a negative category, metrics classified as bias transforming will be less homogeneous than those classified as bias preserving.

¹²⁴ Hardt, Price, and Srebro, *supra* note 121.

¹²⁵ *Id.*

¹²⁶ Chouldechova, *supra* note 12.

¹²⁷ A full taxonomy of fairness metrics along with references is provided in Appendix 1, Table 1a.

¹²⁸ Hardt, Price, and Srebro, *supra* note 121. 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.

¹²⁹ WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 41–44.

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, 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.¹³⁰ 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 (see: Table 1).¹³¹

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.

¹³⁰ In behavioural 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 JOURNAL OF RISK AND UNCERTAINTY 7–59 (1988).

¹³¹ Sahil Verma & Julia Rubin, *Fairness definitions explained*, in 2018 IEEE/ACM INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS (FAIRWARE) 1–7 (2018); 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 PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 514–524 (2020); Sorelle A. Friedler, Carlos Scheidegger & Suresh Venkatasubramanian, *On the (im) possibility of fairness*, ARXIV PREPRINT ARXIV:1609.07236 (2016), <https://arxiv.org/abs/1609.07236> (last visited Nov 11, 2016); Philipp Hacker & Emil Wiedemann, *A continuous framework for fairness*, ARXIV PREPRINT ARXIV:1712.07924 (2017); Alice Xiang & Inioluwa Deborah Raji, *On the Legal Compatibility of Fairness Definitions*, ARXIV PREPRINT ARXIV:1912.00761 (2019); Martim Brandão et al., *Fair navigation planning: A resource for characterizing and designing fairness in mobile robots*, ARTIFICIAL INTELLIGENCE 103259 (2020); Sorelle A. Friedler et al., *A comparative study of fairness-enhancing interventions in machine learning*, in PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 329–338 (2019); Deborah Hellman, *MEASURING ALGORITHMIC FAIRNESS*, 106 VIRGINIA LAW REVIEW 56 (2020).

Fairness metric	Bias preserving?
1. Group fairness, Statistical (demographic) parity	X
2. Conditional statistical (demographic) parity, Conditional independence	X
3. Predictive parity, outcome test	✓
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	X
17. Counterfactual fairness	X
18. No unresolved discrimination	X
19. No proxy discrimination	X
20. Path based causal reasoning	X

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.¹³² 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 metric is essentially a decision about how properties of the target variable distribution should be preserved by a new classifier. As a result, a perfect

¹³² Chouldechova, *supra* note 12; Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent trade-offs in the fair determination of risk scores*, ARXIV PREPRINT ARXIV:1609.05807 (2016); Geoff Pleiss et al., *On fairness and calibration*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 5680–5689 (2017); Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 JOURNAL OF LEGAL ANALYSIS 113–174 (2018).

classifier that preserves all properties of the target distribution with zero error satisfies all of these metrics.

3.3 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.¹³³ 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.¹³⁴

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 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

¹³³ Matt J. Kusner et al., *Counterfactual Fairness* (2017); Niki Kilbertus et al., *Avoiding Discrimination through Causal Reasoning*, ARXIV PREPRINT ARXIV:1706.02744 (2017); Silvia Chiappa & Thomas P. S. Gillam, *Path-Specific Counterfactual Fairness*, ARXIV:1802.08139 [STAT] (2018), <http://arxiv.org/abs/1802.08139> (last visited Jan 16, 2021).

¹³⁴ Chiappa and Gillam, *supra* note 133.

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.¹³⁵ 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 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.¹³⁶ Inheriting social bias in this way cannot be detected by naïve use of bias preserving metrics that simply measure if the machine learning system recovers values of the proxy variables with similar errors for each group.

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 favourable variables to condition on. Nonetheless these choices would, if published alongside summary statistics (which we have advocated for elsewhere¹³⁷), be transparent and open to inspection and debate by affected parties and regulators (see: Section 6.2).

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

¹³⁵ Mitchell et al., *supra* note 33 at 4; Schölkopf, *supra* note 41.

¹³⁶ 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: Case 57325/00 ECHR, D.H. and Others v. the Czech Republic, 2007, <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. CASE 57325/00 ECHR, *supra* note 17; For more details see Chopin, Germaine, and Tanczos, *supra* note 17 at 13–18; FISCHER ET AL., *supra* note 17 at 172–173; With regards to the US and Black and Latin people and immigrants, see HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 40, 120–121, 127, 136.

¹³⁷ WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 62–64.

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.

4 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.¹ 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.

Recognising 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 social biases and inequalities in a given decision-making context.¹³⁸ In contrast, bias transforming metrics require a positive normative choice with regards to which biases should be exhibited by the decision-making system. In making this choice any recognised instance of disparity between groups may be seen as potentially discriminatory and in need of legal justification (see: Section 3.2).

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 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 are often marked by

¹³⁸ HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 120–121, 127, 136. Racist hiring practices, for example, may be indistinguishable from inequality rooted in broader societal factors, such as people of colour having fewer educational opportunities and thus being less competitive in the job market. See: Section 4.

prejudice and inequality.¹³⁹ Historical trends in decision-making have led to diminished and unequal access to opportunities and outcomes among certain groups.¹⁴⁰ 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.¹⁴¹ Racial inequality in top jobs is particularly pronounced in the United States and United Kingdom.¹⁴² People of colour can be treated worse in interviews,¹⁴³ can be less associated with higher

¹³⁹ Evidence cited in this section provides a small glimpse of the vast structural racism, sexism, ableism, and 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 necessarily 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 USA, UK, and EU. All categories and labels used for different demographic categories reflect those used in the primary sources cited.

¹⁴⁰ See for example ANGELA Y. DAVIS, WOMEN, RACE, & CLASS (2011).

¹⁴¹ 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 18 at 65, 190, and 205; 30 YEARS OF BRITISH SOCIAL ATTITUDES SELF-REPORTED RACIAL PREJUDICE DATA, , <https://www.bsa.natcen.ac.uk/media/38110/selfreported-racial-prejudice-datafinal.pdf> (last visited Dec 14, 2020); 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> (last visited Dec 14, 2020). 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, *Ipsos MORI Report on Perceptions and Reality Public Attitudes to Immigration* 56 (2014), <https://www.ipsos.com/sites/default/files/publication/1970-01/sri-perceptions-and-reality-immigration-report-2013.pdf> (last visited Dec 17, 2020).

¹⁴² Statistics from 2019 form the US 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, , <https://www.bls.gov/cps/cpsaat11.htm> (last visited Jul 18, 2020). 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, , <https://www.ethnicity-facts-figures.service.gov.uk/work-pay-and-benefits/employment/employment-by-occupation/latest> (last visited Jul 18, 2020) ‘Elementary’ jobs - the lowest skilled category of occupation recorded in the survey - was highest in the Black (16%) and Other White (15%) ethnic groups.

¹⁴³ Carl O. Word, Mark P. Zanna & Joel Cooper, *The nonverbal mediation of self-fulfilling prophecies in interracial interaction*, 10 JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY 109–120 (1974) demonstrate how interviewers treat Black people differently than white people. For example, interviewers sat further away from

paid occupations, can appear less qualified than their white counterparts with the same qualifications,¹⁴⁴ and can receive lower wages on average.¹⁴⁵

Women and gender non-binary people face similar challenges in the job market.¹⁴⁶ Female associated jobs pay less on average and are seen as less important than male associated jobs.¹⁴⁷ Decisions about promotions and hiring are likewise prone to gender bias,¹⁴⁸ with women routinely rated as less competent than male colleagues despite comparable performance.¹⁴⁹ In the US in 2019 only 27.6% of

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, ESHLEMAN, AND VIJAYA, *supra* note 15 at 155–156 citing this work.

¹⁴⁴ HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 153.

¹⁴⁵ 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 JOURNAL OF HUMAN RESOURCES 701–738 (2007).

¹⁴⁶ For an discussion on gender bias, AI and the workplace see Maria De-Arteaga et al., *Bias in bios: A case study of semantic representation bias in a high-stakes setting*, in PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 120–128 (2019). 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 AMERICAN JOURNAL OF SOCIOLOGY 586–626 (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, *The challenge of being transgender on the academic job market (essay)* | *Inside Higher Ed*, <https://www.insidehighered.com/advice/2016/07/15/challenge-being-transgender-academic-job-market-essay> (last visited Jan 8, 2021). A 2014 US report showed that one out of two transgender individuals are sexually assaulted or abused at some point in their lives. See: Sexual Assault: The Numbers | Responding to Transgender Victims of Sexual Assault, , https://ovc.ojp.gov/sites/g/files/xyckuh226/files/pubs/forge/sexual_numbers.html (last visited Jul 31, 2020).

¹⁴⁷ A study from the US showed that as more women take up particular jobs the sector's perceived prestigiousness and average wages decline. See: PEREZ, *supra* note 20. Women also face greater difficulty when negotiating remuneration due to not being seen as 'likeable' in negotiations. See: Bowles et al., 2007 <https://www.wgea.gov.au/data/wgea-research/gender-equitable-recruitment-and-promotion>. 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 SOCIAL PROBLEMS 455–473 (2002); see also HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 160–161 citing this work.

¹⁴⁸ O'NEIL, *supra* note 23 at 106–122. 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).

¹⁴⁹ For example, in a US 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 et al., *Science faculty's subtle gender biases favor male students*, 109 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 16474–16479 (2012); citing this work see SAINI, *supra* note 16 at 5.

chief executives were women.¹⁵⁰ Physical appearance can also influence employer's assessment of the talent and professionalism of employees and applicants.¹⁵¹

Social stigma around disability also manifests in workplace inequality.¹⁵² People with physical, psychosocial, intellectual, and sensory conditions routinely face significant challenges, biases, and inequalities in the workplace,¹⁵³ including a lack of employment, promotion, mentorship, and significantly higher likelihood of dismissal.¹⁵⁴ Mental conditions in particular are viewed more severely than physical conditions.¹⁵⁵ Ableist assumptions result in perfectly capable workers being disfavoured and disadvantaged. Algorithms satisfying bias preserving metrics can transport these ableist assumptions into future decision-making.¹⁵⁶

Criminal justice is likewise heavily impacted by racial bigotry, with people of colour significantly more likely to be stopped and searched, arrested, and receive harsher punishments than others in the US¹⁵⁷ and UK.¹⁵⁸ Lending decisions are also plagued with racial

¹⁵⁰ Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity, *supra* note 142.

¹⁵¹ 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, ESHLEMAN, AND VIJAYA, *supra* note 15 at 150–151.

¹⁵² 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: DISABLEMENT AND HIERARCHIES OF IMPAIRMENT 7 (2019). For an overview of the cause of this stigma see MICHELLE R. NARIO-REDMOND, ABLEISM: THE CAUSES AND CONSEQUENCES OF DISABILITY PREJUDICE (2019).

¹⁵³ HARPUR, *supra* note 152 at 5–6.

¹⁵⁴ *Id.* at 10–12. Having a disability makes an adult 75 to 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: 298 RICHARD BERTHOUD, THE EMPLOYMENT RATES OF DISABLED PEOPLE (2006); HARPUR, *supra* note 152 at 9.

¹⁵⁵ HARPUR, *supra* note 152 at 2, 14–15.

¹⁵⁶ On ideas on how to counter AI bias in the realm of disabilities see Anhong Guo et al., *Toward Fairness in AI for People with Disabilities: A Research Roadmap*, ARXIV PREPRINT ARXIV:1907.02227 (2019); Os Keyes, *Automating autism: Disability, discourse, and Artificial Intelligence*, 1 THE JOURNAL OF SOCIOTECHNICAL CRITIQUE (2020), <https://digitalcommons.odu.edu/sociotechnicalcritique/vol1/iss1/8>.

¹⁵⁷ In the US, African Americans 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. See: David Cole, *Can our shameful prisons be reformed?*, 1 (2009) as opposed to the 1950s, when segregation was still legal, and around 30% of the African-American population was imprisoned.

¹⁵⁸ 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: Black people nine times more likely to face stop and search than white people, , THE GUARDIAN (2020), <http://www.theguardian.com/uk-news/2020/oct/27/black-people-nine-times-more-likely-to-face-stop-and-search-than-white-people> (last visited Nov 15, 2020). 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 (2013), <https://www.release.org.uk/publications/numbers-black->

bias;¹⁵⁹ race-based ‘redlining’ of postal codes, for example, is still practiced in 2020¹⁶⁰ despite being illegal in many countries.¹⁶¹ Healthcare is another troubling area. Participant samples in clinical trials and health studies are routinely biased towards white males.¹⁶² Medical interventions and metrics often fail to account for biological differences between populations,¹⁶³ for example by treating women as “little men.”¹⁶⁴ 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 and being transgender as a type of identity disorder until 2012.¹⁶⁵

Certain people thus face systematic disadvantage in the workplace, lending, education, criminal justice, health, insurance, and other areas is morally or legally problematic. Merit alone is often insufficient for individuals to succeed compared to their peers.¹⁶⁶ 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.¹⁶⁷ This can be problematic; in

and-white-ethnic-disparities-policing-and-prosecution-drug-offences (last visited Dec 17, 2020).

¹⁵⁹ For example, in the US African-Americans, Latinx, and immigrants are the main targets for predatory lenders which contributed to the 2008 housing crisis. See: STIGLITZ, *supra* note 11 at 88.

¹⁶⁰ 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 Dec 16, 2020).

¹⁶¹ In Germany for example, it would have been illegal to solely use postcodes for lending decisions, see Wachter and Mittelstadt, *supra* note 7 at 96–97; for US 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, ESHLEMAN, AND VIJAYA, *supra* note 15 at 110.

¹⁶² PEREZ, *supra* note 20 at 115–116; on how to address bias in the medical setting see Timo Minssen et al., *Regulatory responses to medical machine learning*, JOURNAL OF LAW AND THE BIOSCIENCES (2020); and Mirjam Pot, Wanda Spahl & Barbara Prainsack, *The Gender of Biomedical Data: Challenges for Personalised and Precision Medicine*, 9 SOMATECHNICS 170–187 (2019).

¹⁶³ PEREZ, *supra* note 20 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), time and resource intensive to study women. SAINI, *supra* note 16 at 58.

¹⁶⁴ 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: SAINI, *supra* note 16 at 59, 62; PEREZ, *supra* note 20 at 116.

¹⁶⁵ HALLEY AND ESHLEMAN, *supra* note 146 at 58. Specifically, until the 1970s the Diagnostic and Statistical Manual of Mental Disorders (DSM) classified “homosexuality” as a mental illness.

¹⁶⁶ STEPHEN J. MCNAMEE & ROBERT K. MILLER, *THE MERITOCRACY MYTH* (2009).

¹⁶⁷ Judith Butler refers to it as the bootstrapping argument JUDITH BUTLER, *GENDER TROUBLE: FEMINISM AND THE SUBVERSION OF IDENTITY* (2011). According to

Western societies, factors such as inheritance, luck, unequal opportunity, and discrimination are just as important to success as merit.¹⁶⁸ For example, the best predictor for whether a person will be in poverty as an adult is whether they were born into poverty.¹⁶⁹

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 technologies), the law protects certain groups (e.g. gender, race, sexual orientation, disability) from direct discrimination and indirect discrimination.¹⁷⁰

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.

¹⁶⁸ 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.” The Meritocracy Myth, , <http://www.ncsociology.org/sociationtoday/v21/merit.htm> (last visited Jul 26, 2020). 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 US. ECONOMIC MOBILITY PROJECT’, *Does America Promote Mobility As Well As Other Nations?* 5 2; in general and citing this work see STIGLITZ, *supra* note 11 at 22. The countries surveyed were the USA, UK, France, Germany, Sweden, Italy, Australia, Finland, Denmark and Canada.

¹⁶⁹ EUBANKS, *supra* note 23 at 205. In the US, 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 et al., *American exceptionalism in a new light: a comparison of intergenerational earnings mobility in the Nordic countries, the United Kingdom and the United States*, 17 (2006) in Nordic countries on the other hand only 20% remain in the poorest group as adults, 30% in the UK; citing this work see HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 106 and ; STIGLITZ, *supra* note 11 at 23. Similarly, Mark Huggett, Gustavo Ventura & Amir Yaron, *Sources of lifetime inequality*, 101 AMERICAN ECONOMIC REVIEW 2923–54, 2949 (2011) state that “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.” See also Alan B. Krueger, *The rise and consequences of inequality in the United States*, 12 SPEECH AT THE CENTER FOR AMERICAN PROGRESS, 3 (2012) who explains that “[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. See also Jonathan Chait, *No such thing as equal opportunity*, NEW YORK MAGAZINE, KASIM 14–16 (2011); see also STIGLITZ, *supra* note 11 at 24 citing this work. 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: F.A.Z.-Serie Schneller Schlau: Wenn die Eltern nicht studiert haben, FAZ.NET, <https://www.faz.net/aktuell/wirtschaft/schneller-schlau/sozialer-aufstieg-wenn-die-eltern-nicht-studiert-haben-16960036.html> (last visited Nov 10, 2020).

¹⁷⁰ WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37; Wachter, *supra* note 45.

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 (see: Section 2).

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 to begin to address, the systemic causes of inequality. To combat systemic inequality and achieve substantive equality, private and public actors must play an active role (see: Section 2.3). In this context, choosing to preserve the status quo must be treated as an explicit normative decision that deems the status quo as acceptable.¹⁷¹ If this choice occurs in sectors known to 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 (see: Section 5).

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. It cannot easily be assumed that a particular type of inequality also occurs in other environments. 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

5 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 our hypothetical scenario in court (see: Section 3.1), 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 (see: Section 2.1), under normal circumstances a claimant would need to provide evidence to convince

¹⁷¹ O’NEIL, *supra* note 23; EUBANKS, *supra* note 23.

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-case basis. For the argument for contextual equality in full, and its significance for fair machine learning, see WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37.

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’).

the court that prima facie discrimination exists by way of showing that a “apparently neutral provision, criterion or practice” disproportionately disadvantages a protected group in comparison with other people.

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 metric¹⁷⁴) 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 interest. Contested measures can be justified if there is a legitimate interest, and the means to achieve it are necessary and proportionate.¹⁷⁵ 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.¹⁷⁶ 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

¹⁷⁴ 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.

¹⁷⁵ For an overview of the ECJ’s case law on legally accepted justifications as well as what type of evidence is admissible, see Wachter, *supra* note 45 at 46–54; WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 41–44.

¹⁷⁶ Cynthia Dwork et al., *Fairness Through Awareness*, ARXIV:1104.3913 [CS] (2011), <http://arxiv.org/abs/1104.3913> (last visited Feb 15, 2016).

variable is the least intrusive possible means of achieving a legitimate interest.¹⁷⁷

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 (see: Section 2.3). They cannot force decision-makers to change their behaviours 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.¹⁷⁸ 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.¹⁷⁹

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 grants decision-makers or developers of automated systems an opportunity 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.¹⁸⁰

While the existence and precise requirements of positive duties remains debated (see: Section 2.3), critically investigating historical

¹⁷⁷ 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; hypothetically, formal equality may be valued higher in certain decision-making contexts in which case bias preserving metrics could be considered necessary.

¹⁷⁸ For example, in the case of conditional independence, the choice of bias is made by choosing which variable(s) to condition on.

¹⁷⁹ 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: Section 6.2).

¹⁸⁰ For a discussion of effective usage of bias transforming metrics, specifically Conditional Demographic Disparity, see: WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 54–64.

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.¹⁸¹ 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.¹⁸² 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 (see: Section 2.3).¹⁸³

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.

6 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

¹⁸¹ EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS AND COUNCIL OF EUROPE, HANDBOOK ON EUROPEAN NON-DISCRIMINATION LAW 239 (2018 edition ed. 2018), https://fra.europa.eu/sites/default/files/fra_uploads/1510-fra-case-law-handbook_en.pdf.

¹⁸² *Id.* at 240.

¹⁸³ 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 US non-discrimination law and automated decision-making. For more details on these doctrines see Nachbar, *supra* note 68 at 23–24 and see also page 51 and 55 where he explains that if knowledge of the discriminatory nature of a “facially neutral” practice exists, it could turn disparate impact into (intentional) disparate treatment.

enough to warrant discussion, and ultimately whether indirect discrimination can be justified.¹⁸⁴

In this paper 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 legislators. To advance the adoption of bias transforming metrics in practice, we conclude with several practical and policy recommendations and open questions for future research.

6.1 A checklist for choosing appropriate fairness metrics

We have argued that bias preserving metrics in decision-making can give rise to *prima facie* indirect discrimination under EU non-discrimination law (see: Section 5). 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 legitimate interest that is pursued in a necessary and proportionate manner).¹⁸⁵ 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.

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 paper. 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

¹⁸⁴ For more detail, see WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37; Wachter, *supra* note 45.

¹⁸⁵ For an overview of the ECJ’s case law on legally accepted justifications, see Wachter, *supra* note 45 at 46–54; WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 41–44.

practice. Both bias preserving and transforming metrics are valuable for diagnostic purposes (see: Section 3.3). Substantive decisions are those with impact on individuals falling within the remit of non-discrimination law.

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 (see: Section 4). 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 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

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. In cases where historical inequality does not exist, or known disparity has been deemed legally justified, both bias preserving and transforming metrics can be used.

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.

equality.¹⁸⁶ The legal acceptability of fairness metrics varies according to the emphasis local legal frameworks place on formal or substantive equality. We have argued that 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. Recognising 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 (see: Sections 3.3 and 5).

6.2 Using bias transforming metrics to support substantive equality

We have argued elsewhere that Conditional Demographic Disparity (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.¹⁸⁷ 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.

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

¹⁸⁶ 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 (see: Section 2.2), EU non-discrimination law aims at substantive equality.

¹⁸⁷ WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37.

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.¹⁸⁸

CDD of course has limitations and is not a silver bullet for algorithmic fairness (see: Section 3.3). Choosing the right conditioning variables is a political decision and developers can be inclined to choose favourable conditions. However, if these choices are – as we recommend – published as summary statistics, these conditions are transparent and are open to inspection and rebuttal.¹⁸⁹

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 (see: Section 2.3). 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.¹⁹⁰

6.3 Substantive equality duties in fair machine learning

While it is clear that the rationale and the aim of non-discrimination law is to dismantle inequality,¹⁹¹ 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 (see: Section 2.3).¹⁹² Specifying the requirements of positive equality duties in fair machine learning is an important area for future research.

¹⁸⁸ *Id.*

¹⁸⁹ *Id.* at 62–64.

¹⁹⁰ 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 and Harpur, *supra* note 8 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.

¹⁹¹ De Vos, *supra* note 59.

¹⁹² 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 66.

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? What are the potential trade-offs?
- Should higher education degrees be allowed as mandatory hiring criterion given inequality in access to education?
- Should we rely on recommendation letters and GPA even if we know that they are biased on 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 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 metric and publishing summary statistics, developers, deployers, or users are forced to make and openly justify a normative judgement as to what bias is locally acceptable. Ultimately courts and legislators will need to decide and develop case law and legislation that shows the path for (active or passive) obligations and responsibilities in dismantling inequality, justifying disparity, and promoting substantive equality for different stakeholders in machine learning and AI.

While bias transforming metrics cannot directly provide a definitive answer regarding the existence and requirements of such duties, their usage creates opportunities for incremental change. This is an important shift away from using machine learning 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.

6.4 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 colour, 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.

Inequality has many and diverse faces. One form of bigotry (e.g. in the US) cannot be assumed to also exist or manifest itself in the same way elsewhere (e.g. Germany). More data is needed to investigate multifaceted inequality globally to promote what we have coined elsewhere as “contextual equality.”¹⁹³ (see: Section 4).

These gaps 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.¹⁹⁴ Failure to collect this data will not prevent discrimination against protected groups, but perhaps make it more difficult to detect.¹⁹⁵ 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).¹⁹⁶

Naturally, privacy scholars have urged to be mindful of the privacy implications of such privacy invasive data collection.¹⁹⁷ This is a legitimate concern and closely related to troubling historical experiences that have significantly harmed minority and marginalised 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

¹⁹³ For more details on “contextual equality”, see WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37.

¹⁹⁴ On the practical limitations of data collection in the EU see TIMO MAKKONEN, MEASURING DISCRIMINATION DATA COLLECTION AND EU EQUALITY LAW (2007).

¹⁹⁵ WACHTER, MITTELSTADT, AND RUSSELL, *supra* note 37 at 34–35; Cynthia Dwork & Deirdre K. Mulligan, *It's not privacy, and it's not fair*, 66 STAN. L. REV. ONLINE 35 (2013); Dwork et al., *supra* note 176; Anupam Datta et al., *Proxy Non-Discrimination in Data-Driven Systems*, ARXIV:1707.08120 [CS] (2017), <http://arxiv.org/abs/1707.08120> (last visited Jan 9, 2021); Kusner et al., *supra* note 133.

¹⁹⁶ Kusner et al., *supra* note 133; Chris Russell et al., *When worlds collide: integrating different counterfactual assumptions in fairness*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 6396–6405 (2017).

¹⁹⁷ VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK AND THINK (2013); For a US 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 63 LEE A. BYGRAVE, DATA PRIVACY LAW: AN INTERNATIONAL PERSPECTIVE (2014); For an European view Sandra Wachter, *Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR*, 34 COMPUTER LAW & SECURITY REVIEW 436–449 (2018); Sandra Wachter, *The GDPR and the Internet of Things: a three-step transparency model*, 10 LAW, INNOVATION AND TECHNOLOGY 266–294 (2018); for a EU and German view see Mario Martini, Wiebke Fröhlich & Saskia Fritzsche, *Algorithmen als Herausforderung für die Rechtsordnung* (2017); for empirical evidence of mobile data collection see Reuben Binns 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 May 11, 2019).

was seen as a step forward.¹⁹⁸ But scientific research, data collection, and databases also contributed to eugenics in Europe, the UK¹⁹⁹ and the US,²⁰⁰ genocide during WWII, racist immigration practices and the denial of basic human rights in the US,²⁰¹ justification of slavery,²⁰² forced sterilisation in the UK,²⁰³ US, Germany and Puerto Rico from the early to the mid-20th Century,²⁰⁴ punishment, castration and imprisonment of LGBT members,²⁰⁵ and denial of equal rights and protection (e.g. sexual violence) to women.²⁰⁶ Privacy²⁰⁷ and harms from slippery slopes in data collection²⁰⁸ 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.²⁰⁹ Pay gaps based on gender and race are painful examples of this reality.²¹⁰ In 2016 in the UK, for example, there was an 18 to 23% gap in wages between men and women (depending on the sector).²¹¹ Gender discrepancies are nothing new, and yet extensive knowledge of them has not yet led to their elimination around the world.²¹²

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

¹⁹⁸ HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 9.

¹⁹⁹ This happened until the 1930's, see EDDO-LODGE, *supra* note 18 at 20–21.

²⁰⁰ HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 36.

²⁰¹ *Id.* at 25.

²⁰² *Id.* at 36–37.

²⁰³ EDDO-LODGE, *supra* note 18 at 20–21.

²⁰⁴ HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 15 at 36–38.

²⁰⁵ HALLEY AND ESHLEMAN, *supra* note 146 at 15–17.

²⁰⁶ SAINI, *supra* note 16 at 233–235.

²⁰⁷ On how data will follow us forever, see VIKTOR MAYER-SCHÖNBERGER, *DELETE: THE VIRTUE OF FORGETTING IN THE DIGITAL AGE* (2011).

²⁰⁸ For surveillance and chilling effects, see JON PENNEY, *Chilling Effects: Online Surveillance and Wikipedia Use* (2016), <https://papers.ssrn.com/abstract=2769645> (last visited Dec 27, 2017).

²⁰⁹ EDDO-LODGE, *supra* note 18 at 208.

²¹⁰ FREDMAN, EUROPEAN COMMISSION, AND EUROPEAN NETWORK OF LEGAL EXPERTS IN THE FIELD OF GENDER EQUALITY, *supra* note 90 at 6.

²¹¹ See, SAINI, *supra* note 16 at 6–7 citing statistic from 2016 .

²¹² 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 best performing with countries leading the way such as Iceland (87.7), Norway (84.2), Finland (83.2 in Sweden (82.0) leading the way.²¹² However, in terms of economic participation and opportunity Western Europe lags behind (69.3%) other players such as North Africa (75.6%) and Eastern Europe and Central Asia (72.2%). North America has closed 73% of their pay gap, sub-Saharan Africa has closed 68% and South Asia two thirds. See: The World Economic Forum, *Global Gender Gap Report 2020*, http://www3.weforum.org/docs/WEF_GGGR_2020.pdf (last visited Aug 28, 2020); PEREZ, *supra* note 20 at 75–78.

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 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 in a 2018 ‘state of the art’ survey paper by Verma and Rubin.²¹³

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 behaviour 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 race). 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, \dots, x_n for all inputs excluding the protected attribute a . Although the substitution $\hat{y} = y$ transforms \hat{y} into a discrete classifier, it still satisfies the continuous definitions (i.e. definitions 10-13).

Evaluating whether 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.

²¹³ Verma and Rubin, *supra* note 131.

Fairness metrics	Formula	Bias preserving?
1. Group fairness, Statistical (demographic) parity ⁱ	$P(\hat{y} = 1 A = a) = P(\hat{y} = 1 A = a') \forall a, a'$	X
2. Conditional statistical (demographic) parity, Conditional independence ⁱⁱ	$P(\hat{y} = 1 C = c, A = a) = P(\hat{y} = 1 C = c, A = a') \forall c, a, a'$	X
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}	$P(y = i \hat{y} = t, A = a) = P(y = i \hat{y} = t, A = a') \forall i \in \{0,1\}, t \in \mathbb{R} a, a'$	✓
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)	$\hat{y}(x_1, x_2, \dots, x_n, a) = \hat{y}(x_1, x_2, \dots, x_n, a') \forall a, a'$	*
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	X
17. Counterfactual fairness ^{xviii}	$\hat{y}_{A \leftarrow a}(x_1, x_2, \dots, x_n, a) = \hat{y}_{A \leftarrow a'}(x_1, x_2, \dots, x_n, a)$	X
18. No unresolved discrimination ^{xix} (causal variant of 2)	$\hat{y}_{A \leftarrow a, X_k \leftarrow x_k}(x_1, x_2, \dots, x_n, a) = \hat{y}_{A \leftarrow a', X_k \leftarrow x_k}(x_1, x_2, \dots, x_n, a)$	X
19. No proxy discrimination ^{xx}	No simple formula	X
20. Path based causal reasoning ^{xxi}	No simple formula	X

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 et al., *Fairness Through Awareness*, ARXIV:1104.3913 [CS] (2011), <http://arxiv.org/abs/1104.3913> (last visited Feb 15, 2016).
- ⁱⁱ Sam Corbett-Davies et al., *Algorithmic decision making and the cost of fairness*, in PROCEEDINGS OF THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 797–806 (2017); Faisal Kamiran, Indrė Žliobaitė & Toon Calders, *Quantifying explainable discrimination and removing illegal discrimination in automated decision making*, 35 KNOWLEDGE AND INFORMATION SYSTEMS 613–644 (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 1193–1216 (2017); Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153–163 (2017).
- ^{iv} Chouldechova, *supra* note 3; Corbett-Davies et al., *supra* note 2.
- ^v Chouldechova, *supra* note 3.
- ^{vi} Moritz Hardt, Eric Price & Nati Srebro, *Equality of opportunity in supervised learning*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 3315–3323 (2016).
- ^{vii} *Id.*
- ^{viii} Richard Berk et al., *Fairness in criminal justice risk assessments: The state of the art*, SOCIOLOGICAL METHODS & RESEARCH 0049124118782533 (2018).
- ^{ix} *Id.*
- ^x *Id.*
- ^{xi} Chouldechova, *supra* note 3.
- ^{xii} Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent trade-offs in the fair determination of risk scores*, ARXIV PREPRINT ARXIV:1609.05807 (2016).
- ^{xiii} *Id.*
- ^{xiv} *Id.*
- ^{xv} Sainyam Galhotra, Yuriy Brun & Alexandra Meliou, *Fairness testing: testing software for discrimination*, in PROCEEDINGS OF THE 2017 11TH JOINT MEETING ON FOUNDATIONS OF SOFTWARE ENGINEERING 498–510 (2017). Not to be confused with the causal methods 17-20 that make use of structured causal models. See: JUDEA PEARL, CAUSALITY (2009).
- ^{xvi} Dwork et al., *supra* note 1.
- ^{xvii} *Id.*
- ^{xviii} Matt J. Kusner et al., *Counterfactual Fairness* (2017).
- ^{xix} Niki Kilbertus et al., *Avoiding Discrimination through Causal Reasoning*, ARXIV PREPRINT ARXIV:1706.02744 (2017).
- ^{xx} *Id.*
- ^{xxi} Razieh Nabi & Ilya Shpitser, *Fair inference on outcomes*, 32 in PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE (2018); Silvia Chiappa & Thomas P. S. Gillam, *Path-Specific Counterfactual Fairness*, ARXIV:1802.08139 [STAT] (2018), <http://arxiv.org/abs/1802.08139> (last visited Jan 16, 2021).