Michigan Technology Law Review

Article 3

2024

The Unfairness of Fair Machine Learning: Leveling Down and Strict Egalitarianism by Default

Brent Mittelstadt
University of Oxford

Sandra Wachter
University of Oxford

Chris Russell
University of Oxford

Recommended Citation
Brent Mittelstadt, Sandra Wachter & Chris Russell, *The Unfairness of Fair Machine Learning: Leveling Down and Strict Egalitarianism by Default*, 30 MICH. TECH. L. REV. (2024).
Available at: https://repository.law.umich.edu/mtlr/vol30/iss1/3

The Unfairness of Fair Machine Learning: Leveling Down and Strict Egalitarianism by Default

Brent Mittelstadt,[1*] Sandra Wachter,[2**] Chris Russell[3***]

ABSTRACT

In recent years, fairness in machine learning (ML), artificial intelligence (AI), and algorithmic decision-making systems has emerged as a highly active area of research and development. To date, most measures and methods to mitigate bias and improve fairness in algorithmic systems have been built in isolation from policymaking and civil societal contexts and lack serious engagement with philosophical, political, legal, and economic theories of equality and distributive justice. Many current measures define "fairness" in simple terms to mean narrowing gaps in performance or outcomes between demographic groups while preserving as much of the original system's accuracy as possible. This oversimplified translation of the complex socio-legal concept of equality into fairness measures is troubling. Many current fairness measures suffer from both fairness and performance degradation—or "leveling down"—where fairness is achieved by making every group worse off or by bringing better-performing groups down to the level of the worst off. Leveling down is a symptom of the decision to measure fairness solely in terms of equality, or disparity between groups in performance and outcomes, that ignores other relevant concerns of distributive justice (e.g., welfare or priority), which are more difficult to quantify and measure. When fairness can only be measured in terms of distribution of performance or outcomes, corrective actions can likewise only target how these goods are distributed between groups. We refer to this trend as "strict egalitarianism by default."

Strict egalitarianism by default runs counter to both the stated objectives of fairness measures and the presumptive aim of the field: to improve outcomes for historically disadvantaged or marginalized groups. When fairness can only be achieved by making everyone worse off in material or relational terms–through injuries of stigma, loss of

[1*] Corresponding author: brent.mittelstadt@oii.ox.ac.uk. Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, United Kingdom.

[2**] Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, United Kingdom.

[3***] Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, United Kingdom. Chris Russell was also an employee of Amazon Web Services during part of the writing of this article. He did not contribute to this research in his capacity as an Amazon employee. This work has been supported through research funding provided by the Wellcome Trust (grant nr 223765/Z/21/Z), Sloan Foundation (grant nr G-2021-16779), the Department of Health and Social Care (via the AI Lab at NHSx), and Luminate Group to support the Trustworthiness Auditing for AI project and Governance of Emerging Technologies research program at the Oxford Internet Institute, University of Oxford.

solidarity, unequal concern, and missed opportunities for substantive equality—something has gone wrong in translating the vague concept of "fairness" into practice. Leveling down should be rejected in fairML because it (1) unnecessarily and arbitrarily harms advantaged groups in cases where performance is intrinsically valuable, such as medical applications of AI; (2) demonstrates a lack of equal concern for affected groups, undermines social solidarity, and contributes to stigmatization; (3) fails to live up to the substantive aims of equality law and fairML and squanders the opportunity afforded by interest in algorithmic fairness to substantively address longstanding social inequalities; and (4) fails to meet the aims of many viable theories of distributive justice including pluralist egalitarian approaches, prioritarianism, sufficientarianism, and others.

This paper critically scrutinizes these initial observations to determine how fairML can move beyond mere leveling down and strict egalitarianism by default. We examine the causes and prevalence of leveling down across fairML and explore possible justifications and criticisms based on philosophical and legal theories of equality and distributive justice as well as equality-law jurisprudence. We find that fairML does not currently engage in the type of measurement, reporting, or analysis necessary to justify leveling down in practice. The types of decisions for which ML and AI are currently used, as well as inherent limitations on data collection and measurement, suggest leveling down is rarely justified in practice. We propose a first step toward substantive equality in fairML: "leveling up" systems by enforcing minimum acceptable harm thresholds, or "minimum rate constraints," as fairness constraints at the design stage. We likewise propose an alternative harms-based framework to counter the oversimplified egalitarian framing currently dominant in the field and push future discussion more towards substantive equality opportunities and away from strict egalitarianism by default.

INTRODUCTION[4****]

In recent years fairness in machine learning (ML), artificial intelligence (AI), and algorithmic decision-making systems has emerged as a highly active area of research and development. Predicted and actual uses of these technologies to distribute outcomes and resources in high-stakes domains such as medicine, law, and finance have rightly driven interest in the fairness and bias of those decisions. Deployment of these technologies has been contested by policymakers[5] and researchers[6] based on a perceived lack of trustworthiness and fairness.

The field of fair machine learning (fairML) has been predominantly driven by researchers and practitioners working in ML, AI, computer science, software engineering, and mathematics. These groups have developed numerous measures and methods to mitigate bias and improve fairness in algorithmic systems. However, the majority of these tools have been built in isolation from policymaking and civil societal contexts and lack serious engagement with philosophical, political, legal, and economic theories of equality and distributive justice.[7] Reflecting this, most define "fairness" in simple terms to mean narrowing gaps in performance or outcomes between demographic groups. Successfully achieving algorithmic fairness has come to mean satisfying one of these simple mathematical definitions, while preserving as much of the accuracy of the original system as possible.[8]

This oversimplification of equality through fairness measures could be attributed to the relative youth of fairML. However, the practical impact of the approach adopted by the field to date is morally troubling. Many current fairness measures have been shown to suffer from both

[4****] This paper would not exist without Jade Thompson, Professor Andrew Przybylski, and our amazing colleagues in the Governance of Emerging Technologies research program at the Oxford Internet Institute. This work has been supported through research funding provided by the Wellcome Trust (grant nr 223765/Z/21/Z), Sloan Foundation (grant nr G-2021-16779), the Department of Health and Social Care (via the AI Lab at NHSx), and Luminate Group.

[5] E.g., GOVERNMENT OFFICE FOR SCIENCE, ARTIFICIAL INTELLIGENCE: OPPORTUNITIES AND IMPLICATIONS FOR THE FUTURE OF DECISION MAKING, (2016) (UK) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf; HOUSE OF COMMONS SCIENCE AND TECHNOLOGY COMMITTEE, THE BIG DATA DILEMMA, 52, HC 468, at 52 (UK) (2016) https://publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf.

[6] Nina Grgic-Hlaca et al., *Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Predictions*, PROC. 2018 WORLD WIDE WEB CONF. 903 (2018); Maximilian Kasy & Rediet Abebe, *Fairness, Equality, and Power in Algorithmic Decision-Making*, PROC. 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 576 (2021).

[7] Reuben Binns, *On the Apparent Conflict Between Individual and Group Fairness*, PROC. 2020 CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 514, 515 (2020).

[8] *See* Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, PROC. 2019 CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 59, 61–63 (2019).

fairness and performance degradation, or "leveling down," where fairness is achieved by making every group worse off, or by bringing better-performing groups (i.e., "advantaged groups") down to the level of worse-performing groups (i.e., "disadvantaged groups").[9] Leveling down is effectively fairness achieved by breaking the system, for example, by making a classifier less accurate so it performs equally badly across all relevant groups.

Leveling down is a symptom of the decision to measure fairness solely in terms of equality, or disparity between groups in performance and outcomes,[10] while ignoring other relevant features of distributive justice such as absolute welfare or priority, which are more difficult to quantify and directly measure in research and development environments[11]. When fairness can only be measured in terms of distribution of performance or outcomes, corrective actions can likewise only target how these goods are distributed between groups. The field effectively only has egalitarian tools at its disposal that value equality of treatment and outcomes while ignoring other goods of distributive justice. Likewise, the prevalence of leveling down in fairML suggests that the field is, intentionally or otherwise, adopting a strict egalitarian approach to questions of distributive justice in which the only (measurable) value is equality. We name these trends in fairML "strict egalitarianism by default."

Strict egalitarianism by default, at least in its most gratuitous forms, runs counter to both the stated objectives of fairness measures and the presumptive aim of the field: to improve outcomes for historically disadvantaged or marginalized groups.[12] It conceives of equality in

[9] Dominik Zietlow et al., *Leveling Down in Computer Vision: Pareto Inefficiencies in Fair Deep Classifiers*, arXiv (2022), https://doi.org/10.48550/arXiv.2203.04913.

[10] *See infra* Table 15; The exclusive focus on equality defined as disparity between groups is a basic feature of "group fairness" measures. *See also* Verma & Rubin *infra* note 13.

[11] For further discussion on welfare and priority, *see infra* Section III.A. and III.A.1.

[12] FairML does not have universally agreed guiding principles, but prior work can provide some indication of its values and aims. In a 2019 paper critical of the state of the field, Keyes et al. defined the 'Fair' value of the Fairness, Accountability, and Transparency in Machine Learning (FAT-ML or FAccT-ML) research network as ensuring that algorithmic systems are "lacking biases which create unfair and discriminatory outcomes." *See* Os Keyes, Jevan Hutson & Meredith Durbin, *A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry*, EXTENDED ABSTRACTS 2019 CHI CONF. ON HUMAN FACTORS IN COMPUTING SYSTEMS 1 (2019), https://doi.org/10.1145/3290607.3310433. More recently, Cooper et al. suggest that the field is motivated by the fact that "automated decision systems that do not account for systemic discrimination in training data end up magnifying that discrimination; to avoid this, such systems need to be proactive about being fair." *See* A. Feder Cooper et al., *Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research*, PROC. 2021 AAAI/ACM CONF. ON AI, ETHICS, AND SOCIETY 46, 51 (2021) https://dl.acm.org/doi/10.1145/3461702.3462519. Early case studies in the field are similarly instructive, such as the famous COMPAS case in which a risk recidivism algorithm was alleged by journalists at ProPublica to be biased against Black defendants, routinely assigning them higher risk scores than comparable white defendants. *See* Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016),

simplistic comparative terms, ignoring the absolutes of welfare and justice that are necessary to achieve substantive equality rather than mere formalistic equality, or equal treatment.[13]

When fairness can only be achieved by making everyone worse off in material or relational terms—through injuries of stigma, loss of solidarity, unequal concern, and missed opportunities for substantive equality—something would appear to have gone wrong in translating the vague concept of "fairness" into practice. Equality should aim to make people better off, not reduce them to a common level of harm.[14] Simple mathematical definitions can be satisfied without regard for how parity is achieved in practice including the significant material and relational harms and opportunity costs for the people affected.

The huge interest that exists in algorithmic fairness provides an opportunity to substantively address longstanding inequalities in society. Enforcing fairness solely through leveling down squanders this chance to achieve substantive rather than mere formalistic equality.

This paper scrutinizes these initial observations to determine whether, and to what extent, leveling down and strict egalitarianism by default are problematic for fairML. Sections I and II introduce the concept of leveling down and examine its prevalence across fairML research and development. Section III draws on philosophical and legal theories of equality and distributive justice, as well as equality law jurisprudence, to explore possible justifications and criticisms of leveling down as a tool of distributive justice. Section IV then considers the relevance and feasibility of these possible justifications in the context of fairML, concluding that the field does not currently engage in the type of thinking necessary to justify equality achieved through leveling down. The types of decisions for which ML and AI are currently used, as well as inherent limitations on data collection and measurement, both suggest leveling down is rarely justified in practice. Section V describes an alternative approach to fairness by "leveling up" systems by enforcing minimum acceptable harm thresholds, or "minimum rate constraints," as fairness constraints at the design stage. We propose an alternative

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. These examples suggest work on algorithmic fairness is motivated at least in part by a desire to improve the situation of disadvantaged people that unjustifiably receive worse treatment or outcomes than their peers. How fairness, discrimination, and bias are conceptualized and measures, and likewise what is justified in differential treatment, opportunities, and results of course differs drastically across the field and use cases, but the underlying motivation to help people who are unjustifiably harmed by algorithmic systems seems clear and uncontroversial.

[13] Sandra Wachter et al., *Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law*, 123 W. VA. L. REV. 735, 744–745 (2021).

[14] *See generally* LARRY S. TEMKIN, INEQUALITY (Louis P. Pojman & Robert Westmoreland eds., 1993); Nils Holtug, *Egalitarianism and the Levelling Down Objection*, 58 ANALYSIS 166, 166–67 (1998); Brett Doran, *Reconsidering the Levelling-down Objection Against Egalitarianism*, 13 UTILITAS 65, 84 (2001); Derek Parfit, *Equality or Priority?*, *in* THE IDEAL OF EQUALITY 81–2 (Matthew Clayton & Andrew Williams eds., 2002).

harms-based framework to counter the oversimplified egalitarian framing currently dominant in the field and to push future discussion toward substantive equality opportunities. The Conclusion provides recommendations for a normative and practical reconfiguration of the field to resist leveling down and strict egalitarianism by default.

I. LEVELING DOWN

Methods which enforce group-based parity measures of fairness (or "group fairness" measures) are morally and legally problematic due to (1) their implicit values, which disproportionately favor equality over reduction of harm, and (2) the way they achieve equality in practice through "leveling down," by which certain groups are needlessly made worse off for the sake of mathematical convenience. Far from being a mathematical or theoretical exercise, the enforcement of fairness in these terms arbitrarily harms people subject to decisions informed by "fair" ML or AI systems.[15]

We can decompose the machinery of fair ML into two components: measures and methods. "Measures" are simple computations, such as the difference in true positive rate between two protected groups, that describe how unfair the behavior of a system is or appears to be. "Methods" are novel approaches to ML that improve these measures, equalizing rates of harm at the expense of criteria such as overall accuracy, which are typically optimized by an ML system trained without consideration of fairness.

When we build ML systems to make decisions about people's lives, our design decisions encode implicit value judgments about what properties the system should prioritize. For example, standard ML systems are typically trained to maximize some notion of accuracy by minimizing a proxy such as log loss.[16] Methods used to enforce fairness in ML systems, or "algorithmic fairness methods," likewise impose certain value judgements about the properties a system should optimize, for example valuing equality of error rates over accuracy, and alter system behavior accordingly.

The idea that accuracy is not always the most relevant property for evaluating a model's performance is commonly accepted across ML research. For example, when dealing with rare events such as trying to identify forms of cancer that occur in less than 1% of the population, a constant classifier that always predicts cancer is not present will have over 99% accuracy for any representative sample. It may likewise have

[15] For example, in the case of hiring, lower hiring rates, or in the case of cancer detection, an increased failure to identify people who have cancer as having cancer.

[16] *See generally* VLADIMIR N. VAPNIK, THE NATURE OF STATISTICAL LEARNING THEORY (Michael Jordan et al. eds., 2nd ed. 1999).

higher accuracy than other models or classification methods that would, nonetheless, be more useful in practice.

In such cases involving severely unbalanced datasets, properties such as precision (e.g., the proportion of people predicted as being at-risk of cancer who actually have cancer) or recall (e.g., the proportion of the people who will eventually develop cancer who are correctly predicted as at-risk of cancer) are often more useful. In the case of rare cancer screening, positive decision rates (e.g., the proportion of people predicted as at-risk) may be a more relevant property to optimize. For example, a healthcare authority may only have resources to screen at most *k%* of the population, in which case they may prefer a model that maximizes recall while keeping the percentage of the population called for screening under this *k%*.

A. Leveling Down via Group Fairness

Standard group-based parity measures of fairness, or "group fairness," tend to achieve fairness by selecting one or more properties that are more important than accuracy for a particular case, and then enforcing equality for this property across relevant demographic groups while preserving accuracy as far as possible.[17] Example measures include equality of opportunity (corresponding to equality of recall across demographic groups), equality of precision, and demographic parity (corresponding to equality of positive rate).[18]

Once a property has been selected, equality can be enforced in two ways: (1) adjust performance along the chosen property for the disadvantaged group(s), for example by improving recall at the cost of accuracy, and (2) degrade performance for advantaged group(s) along the same property. In practice, these approaches tend to be combined to satisfy group fairness measures as fully as possible.

Concerning the former, enforcing equality for one or more properties while also maximizing accuracy often requires altering the behavior of a classifier for multiple groups.[19] For example, groups which have a below average positive decision rate or recall, henceforth referred to as "disadvantaged groups," can have these properties increased by enforcing a group fairness method on the classifier. Gains in positive decision rates or recall typically come at the cost of accuracy.

[17] *See* Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, PROC. 23RD ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 797 (2017), https://doi.org/10.48550/arXiv.1701.08230; Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, 2018 ACM/IEEE INT'L WORKSHOP ON SOFTWARE FAIRNESS (FAIRWARE) 1, 3 (2018).

[18] Verma and Rubin, *supra* note 13.

[19] Corbett-Davies et al., *supra* note 13.

Nonetheless, if groups are thought to be harmed by low decision rates or low recall, this trade-off can be considered beneficial or justified.

Concerning the latter, group fairness measures also tend to alter performance for groups with above average performance, henceforth referred to as "advantaged groups." Performance for relevant properties, such as decision rate or recall, tend to be degraded. This degradation likewise comes at the cost of accuracy.[20] Assuming that performance measures like higher decision rates or recall are inherently valuable properties, and higher accuracy is likewise valuable, the resulting fair classifier would be Pareto inefficient.[21]

Enforcing equality by degrading performance for advantaged groups causes the phenomenon we refer to as "leveling down" Performance is arbitrarily degraded for advantaged group(s) solely to reduce disparity in a given property (e.g., recall, decision rates) between groups while maintaining as much accuracy as possible. Leveling down occurs where equality is enforced not only by increasing the relevant property for disadvantaged groups, but by arbitrarily making one or more better-performing groups worse off by reducing performance for them along the same property.

Here, we are concerned with cases where the degradation of performance for advantaged groups is not causally linked to improvements in performance for disadvantaged groups. In such cases, the loss of performance for advantaged groups is not strictly necessary to improve recall, decision rates, or some other valuable property for disadvantaged groups; rather, it is inflicted solely to reduce performance disparity in the chosen property, thereby satisfying a mathematical definition of fairness which says a model is fair when the performance value is equal between groups.

Our concern arises because this behavior introduces additional and avoidable harm to advantaged groups solely to achieve parity between groups. Leveling down does not directly benefit disadvantaged groups. A more instructive framing to capture this behavior is to think of the choice of group fairness measures as a choice about the type of harm that should be equalized among groups.[22]

Equal opportunity, for example, requires recall rates to be equalized between groups while maintaining as much accuracy as possible. The harm caused is a loss of recall for advantaged groups, or a greater failure

[20] *See id.*

[21] Zietlow et al., *supra* note 5, at 1. This inefficiency means that for some groups, both the performance metric on these groups (decision rate or recall) and classifier accuracy on the group are needlessly decreased.

[22] A table examining the type of harm equalized by different group fairness measures is available in Appendix 1. *Infra* Figure 1. In no small part, the wide range of fairness measures now available in fairML can be attributed to researchers identifying relevant properties from the classification literature and then finding ways to enforce equality with respect to these properties. *See* Verma & Rubin, *supra* note 13.

to correctly identify positive cases. A mathematically optimal solution that equalizes recall at a minimum loss to accuracy would involve both steps mentioned above: increase recall for groups that were disadvantaged by the original classifier but also reduce it for previously advantaged groups.[23] If enforced on a cancer screening system, for example, equalizing recall means that more cases of cancer will be missed for advantaged groups than would have otherwise been the case. What is not clear, and what we will examine in this paper, is whether the avoidable harms caused by leveling down for both advantaged and disadvantaged groups can be ethically, legally, and socially justifiable.

B. Example: Leveling Down in Cancer Screening

As an illustrative example of the harm of leveling down in practice, imagine that we want to enforce fairness in an AI system used for predicting future risk of lung cancer. Our hypothetical system, inspired by a real-world patient triage system,[24] suffers from a performance gap between Black and white patients. Specifically, the system has lower recall for Black patients, meaning it routinely underestimates their risk of cancer and incorrectly classifies patients who will eventually develop lung cancer as "low risk."

The inferior baseline performance can have many causes. It may have resulted from training our system on data predominantly taken from white patients, or because health records from Black patients are less accessible or of a lower quality. Likewise, it may reflect underlying worse performance in existing clinical technologies, or social inequalities in healthcare access and expenditures. For example, during the COVID-19 pandemic, pulse oximeters overestimated blood oxygen levels in minorities resulting in delayed treatment for darker-skinned patients.[25] Similarly, lung and skin cancer detection technologies have been shown to be less accurate for darker-skinned people, resulting in an increased failure to flag cancers in patients, delaying access to life-saving care.[26] Furthermore, patient-triage systems regularly underestimate the need for care in minority ethnic patients. For example, such systems use health care costs as a proxy for illness while failing to account for unequal access to care, and thus unequal costs, across

[23] Corbett-Davies et al., *supra* note 13, at 801 n.8.

[24] Ziad Obermeyer & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm That Guides Health Decisions for 70 Million People*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 89 (Jan. 19, 2019) (citing abstract).

[25] Ashraf Fawzy et al., *Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19*, 182 JAMA INTERNAL MED. 730, 734–35 (2022).

[26] David Wen et al., *Characteristics of Publicly Available Skin Cancer Image Datasets: A Systematic Review*, 4 LANCET DIGITAL HEALTH e64, e70–71 (2022).

populations.[27] Whatever the cause of the performance gap, our motivation for enforcing fairness is to substantively improve performance for a worse-performing (or “disadvantaged”) group, in this case Black patients.

In the context of cancer screening, false negatives are much more harmful than false positives; the latter mean that the patient will have unnecessary health checks or scans, whereas the former mean that future cases of cancer will go undiagnosed and untreated.

One way to improve the situation for Black patients is therefore to improve the system’s recall. As a first step, we may decide to err on the side of caution and program the system to change its predictions for the cases involving Black patients that it is least confident about. Specifically, we would flip some low confidence “low risk” cases to “high risk” to catch more cases of cancer in the future. This change comes at the cost of accuracy; the number of people incorrectly classified as being at risk of cancer goes up, and the system’s overall accuracy goes down. However, this trade-off between accuracy and recall is often seen as acceptable on the basis that failing to predict a future cancer case severely harms patients more than the additional, unnecessary screening caused by a false positive.

By flipping cases in this way to increase recall at the cost of accuracy we can eventually reach a state where any further changes would come at an unacceptably high loss of accuracy. Where this threshold lies in practice is ultimately a subjective decision; there is no objective tipping point between recall and accuracy. We have not necessarily brought performance (or recall) for Black patients up to the same level as white patients, but we have done as much as possible within the constraints of the current system and available data and other resources to improve the situation of Black patients and reduce the performance gap.

Here, we encounter the key dilemma responsible for leveling down in fairML. We can take an optional second step to further reduce the performance gap between Black and white patients. We cannot improve performance for Black patients any further without an unacceptable loss of accuracy. However, we can still reduce performance for white patients, lowering both their recall and accuracy in the process, so that our system performs equally well, or as close as possible, for both groups.

[27] Obermeyer & Mullainathan, *supra* note 20. Similar bias can also be observed along gender lines, with female patients disproportionately misdiagnosed and mistreated for heart disease. *See* Nancy N. Maserejian et al., *Disparities in Physicians’ Interpretations of Heart Disease Symptoms by Patient Gender: Results of a Video Vignette Factorial Experiment*, 18 J. WOMEN'S HEALTH (LARCHMT) 1661 (2009); Linda Worrall-Carter et al., *Systematic Review of Cardiovascular Disease in Women: Assessing the Risk*, 13 NURSING & HEALTH SCIENCES 529 (2011).

By definition, this is what many group fairness methods do in practice.[28] The motivation is mathematical convenience: the aim is to make two values (e.g., recall) as close to equal as possible between two groups (e.g., white and Black patients), solely to satisfy a mathematical definition that characterizes a system as fair when these two numbers are equal.[29] In our example, leveling down would mean altering the labels of white patients as well, switching some of the predictions from high- to low-risk. Clearly, this type of label flipping for the sake of equality can be extremely harmful for patients who would not be offered follow-up care and monitoring. Overall accuracy decreases and the frequency of the most harmful type of error increases, all for the sake of reducing the performance gap. In our example, this leveling down does not improve the situation of Black patients (who already have a classifier with improved recall); rather, it serves only to equal out performance (or recall) between Black and white patients. It can moreover cause broader social harms, undermine more difficult but substantively rich solutions to inequality (e.g., increased access to healthcare, improved data quality), and engender stigmatization and social isolation.[30]

Leveling down thus benefits neither group directly. Assuming a system has already been designed to minimize costs for all patients, it would be inappropriate to choose an intervention that "would inevitably make at least one group worse off without making the other group better off."[31] Yet, this is precisely what current applications of group fairness achieve in fairML.

II. HOW COMMON IS LEVELING DOWN IN FAIRML?

The use of equality-based fairness measures in machine learning is not self-evidently troubling; rather, it is the act of enforcing these definitions algorithmically,[32] and the resulting leveling down, which causes

[28] Zietlow et al., *supra* note 5; Lily Hu & Yiling Chen, *Welfare and Distributional Impacts of Fair Classification*, ARXIV (2018), http://arxiv.org/abs/1807.01134 (last visited July 13, 2022); s *ee also infra,* Section II.

[29] *See infra*, Section II.

[30] *See infra* Section IV.

[31] Chloé Bakalar et al., *Fairness on the Ground: Applying Algorithmic Fairness Approaches to Production Systems*, ARXIV 1, 5 (2021).

[32] Examples of algorithmic enforcement strategies include: postprocessing, *see* Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, 29 ADVANCES IN NEURAL INFO. PROCESSING SYS. 3315 (2016), Corbett-Davies et al., *supra* note 13; retraining, *see* Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification Without Discrimination*, 33 KNOWLEDGE AND INFO. SYS. 1 (2012), Alekh Agarwal et al., *A Reductions Approach to Fair Classification*, 80 INT'L CONF. ON MACH. LEARNING 60 (2018); novel objectives, *see* Muhammad B. Zafar et al., *Fairness Constraints: Mechanisms for Fair Classification*, 54 A.I. & STAT. 962 (2017), Michael Lohaus et al., *Too Relaxed to be Fair*, 119 INT'L CONF. ON MACH. LEARNING 6360 (2020).

problems. When used solely for diagnostic purposes, egalitarian measures such as (conditional) demographic parity[33] or equal opportunity[34] provide a helpful warning that different groups are being treated differently, and that they are potentially harmed in different ways by an algorithmic decision-making system. However, using such egalitarian measures to determine which models should be deployed in real world use cases raises serious ethical and legal concerns. Leveling down can occur as a direct result of the use of egalitarian measures in model selection. In many ways this problem is an example of Goodhart's Law that "when a measure becomes a target, it ceases to be a good measure."[35]

In this section we demonstrate how leveling down occurs for a range of fairness measures, focusing on two of the most common metrics in the fairML literature: demographic parity and equal opportunity. To demonstrate, we enforce these measures across a range of algorithms using two of the most prevalent fairness toolkits: FairLearn and IBM AI Fairness 360 (IBM360).[36] Through this analysis, we show that leveling down is not a limitation or design flaw of these toolkits or methodologies. Rather, it is a natural consequence of strictly enforcing equality in model selection.[37] As such, we show that leveling down is a concern for anyone who takes group fairness measures into consideration when deciding which model should be deployed—not just those who try to enforce fairness through ML toolkits.

A. Why Does Leveling Down Occur?

To understand why leveling down occurs it is vital to first recognize that most notions of group fairness are underspecified. For example, to enforce demographic parity it is only necessary to create a classifier that

[33] Demographic Parity balances the decision rates among groups. Sandra Wachter, Brent Mittelstadt & Chris Russell, *Why Fairness Cannot Be Automated: Bridging the Gap B etween EU Non-Discrimination Law and AI*, 41 COMPUT. L. & SEC. REV. 105567 (2021); Kamiran & Calders, *supra* note 28.

[34] Equal Opportunity balances the recall rate among groups. Hardt, Price & Srebro, *supra* note 28.

[35] Marilyn Strathern, *'Improving Ratings': Audit in the British University System*, 5 EUR. REV. 305, 308 (1997).

[36] Rachel K.E. Bellamy et al., *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*, 63 IBM J. RSCH. & DEV. 4:1 (2019); Sarah Bird et al., Fairlearn: A Toolkit for Assessing and Improving Fairness in AI , MICROSOFT TECH. REP. MSR-TR-2020-32 (2020).

[37] For example, Kim discusses a range of changes to the design of an ML algorithm that could decrease the "disparate impact" (a concept from US anti-discrimination law loosely corresponding to demographic parity) of the decisions made by a system. *See* Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action*, 110 CAL. L. REV. 1539, 1575–83 (2022). If a data scientist systematically explored combinations of these changes, and then selected the model with disparate impact below a predetermined threshold, and that otherwise maximizes accuracy, it is likely that this would also exhibit leveling down.

gives positive decisions to 100% of people, 0% of people, or any ratio in between, provided that the same proportion of each group (e.g., 10% of Black and 10% of white patients) receive positive decisions. The same ambiguity occurs when enforcing equal opportunity, only here the recall of each group must be matched rather than the ratio of positive decisions.[38]

Given this ambiguity, how are models actually selected using group fairness measures? In practice, models are selected to maximize some notion of classification performance—such as accuracy, balanced accuracy, F1-score, or Matthew's Correlation Coefficients (MCC)[39]—while also being sufficiently equal to satisfy a chosen fairness measure as far as possible. Reflecting this, comparisons between models on the basis of fairness and accuracy are common in the ML literature. Methods are competitively benchmarked based on their ability to obtain higher levels of accuracy for a given level of fairness.[40]

It is this combination of maximizing accuracy while enforcing equality that leads to leveling down. To understand why, it is necessary to briefly discuss how ML classifiers operate.

ML classifiers tend not only to assign labels to datapoints corresponding to individuals, but also a notion of confidence indicating how likely it is that a particular label is correct. Low-confidence datapoints (e.g., an individual patient's predicted cancer risk) are least likely to be labeled correctly. Consequently, altering low confidence cases can substantially change an equality-based fairness measure (e.g., recall rate) with little impact on the classifier's overall accuracy. As such, to enforce equality while maximizing accuracy, it is beneficial to alter low confidence cases in both groups because these have the lowest "cost" in terms of accuracy.

This approach brings the groups closer to parity along a given equality measure—for example, recall rate—by increasing performance (or, here, recall) for disadvantaged groups and reducing performance (or recall) for advantaged groups. The alternate strategy of leveling up, and simply increasing performance for disadvantaged groups until they obtain parity with the most advantaged groups, requires labels to be altered for datapoints where the classifier is more confident, resulting in

[38] More generally this ambiguity holds for any notion of group equality. If we want the groups to be (approximately) equal with respect to some property such as selection rate, recall, or precision, this still leaves open the question as to what specific value should the property take.

[39] Yasen Jiao & Pufeng Du, *Performance Measures in Evaluating Machine Learning Based Bioinformatics Predictors for Classifications*, 4 QUANTITATIVE BIOLOGY 320, 325–6 (2016).

[40] Model benchmarking reflects an underlying truth about how ML is used in practice: it is used precisely because some notion of classifier performance is held to be important in a given decision-making process. If this was not the case outcomes could instead be assigned arbitrarily, and there would be no need for ML. This is not to say that performance alone is sufficient. Fairness methods are used in order to maximize performance while satisfying other criteria.

a greater drop in accuracy.[41] This relationship between confidence and accuracy was formalized in a 2017 paper by Corbett-Davies et al. that showed that if a classifier is well-calibrated[42] then a greedy strategy that systematically disadvantages already advantaged groups and advantages already disadvantaged groups is provably optimal.[43]

However, systematically disadvantaging some groups while advantaging others is optimal in a much wider range of scenarios. For example, whenever a classifier uses data about some set of individuals that is intrinsically uninformative, the classifier will exhibit poor accuracy for these and similar individuals. Decisions about these individuals can therefore be altered with minimal loss of accuracy. Disadvantaging these "difficult to label" individuals in advantaged groups, and likewise advantaging difficult to label individuals in disadvantaged groups, allows for substantial reductions of inequality with relatively little loss of accuracy.

While only the post-processing method described by Corbett-Davies et al. explicitly performs this type of "leveling down" to achieve equality between groups,[44] other methods for enforcing fairness in ML implicitly behave the same way. Lohaus et al., for example, found that a range of methods for enforcing demographic parity in deep networks exhibit the same trade-off between equality and accuracy and make statistically indistinguishable decisions about individuals.[45] Lipton et al. find similar results for simple classifiers.[46]

As such, leveling down is often an optimal solution to satisfy a fairness measure while retaining as much accuracy as possible. Enforcing group fairness need not, however, always result in leveling down. There are at least two broad cases in which leveling down need not occur: (1) when model selection fails and (2) when models are insufficiently expressive to treat different groups differently.

Concerning model selection failure, Zietlow et al. analyzed a range of published approaches for bias-preserving fairness in computer vision,[47] observing that none used held-out data to determine error rates.[48] This failure to use held-out data is troubling in computer vision

[41] *See infra* Section V.

[42] A classifier is *well-calibrated* if, for every group, the confidence score it returns corresponds to the probability that the classifier is correct.

[43] Corbett-Davies et al., *supra* note 13.

[44] *Id.*

[45] Michael Lohaus et al., *Are Two Heads the Same as One? Identifying Disparate Treatment in Fair Neural Networks*, ARXIV, 1 (2022).

[46] Zachary C. Lipton, Julian McAuley & Alexandra Chouldechova, *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, 31 ADVANCES IN NEURAL INFO. PROCESSING SYS. (2018).

[47] Zietlow et al., *supra* note 5. Held-out data refers to data that was not used for the training of a machine learning classifier, but that was instead held back to allow for the accurate evaluation of classifier performance.

[48] *Id.*

where, because the usage of high-capacity models means that training error goes to zero, the only way to reliably estimate error rates is by using held-out data. As such, while the reviewed approaches did show a decrease in both accuracy and inequality, it was more likely due to a general deterioration in model performance than because fairer models were explicitly selected.[49]

Concerning inexpressive models, recall the discussion above regarding how fair classifiers that maximize accuracy should behave.[50] There, it is assumed that "difficult to label" individuals can be easily identified, and that their group membership can be inferred. However, in some cases, this may not be true. For example, if classifiers do not have access to data about group membership, it may not be possible to infer group membership reliably enough to differentiate treatment or "flip" labels in an informed way.

In both cases, the behavior is even more concerning than leveling down. In the standard form of leveling down discussed initially we can be confident that at least one disadvantaged group is better off, and that the measure we are trying to equalize (e.g., recall or selection rate) has improved for this group. In cases of model selection failure and inexpressive models, we cannot be confident in this knowledge. While a method for enforcing group fairness must result in less inequality to be considered a success, it may do this by decreasing or increasing performance for every group, and there is no guarantee as to how this will be accomplished in practice. For example, Zietlow et al. found that most fairness methods in computer vision improved equal opportunity by decreasing the average recall for every group across a range of tasks.[51] In many high-risk situations, such as medical testing, the use of methods that improve equality by decreasing performance (e.g., diagnosis rates) for everyone would be grossly inappropriate.

B. Leveling Down in Practice

To further establish the prevalence of leveling down in fairML, we demonstrate its occurrence using two of the most popular fairness toolkits using real-world code bases. Specifically, we illustrate how leveling down occurs using standard examples taken from the "How To" guides for FairLearn and IBM AI Fairness 360.

For FairLearn, we ran the code from the project's quick start guide.[52] This code makes use of the exponentiated gradient algorithm of Agarwal

[49] *Id.*

[50] *See supra* Section II.A.

[51] Zietlow et al., *supra* note 5, at 1.

[52] *Quickstart*, supra.

et al. [53] to enforce fairness using the decision tree implementation of scikit-learn.[54] Results are shown on the UCI Adult Dataset.[55]

We made two minor modifications to the instructions and code provided in the FairLearn quick start guide. First, we ran the code three times, enforcing all of the group fairness metrics supported by the toolkit: demographic parity, True Negative rate, and True Positive Rate, the latter of which corresponds to equal opportunity (or equal recall).[56] Second, we increased the maximum tree depth from 4 to 10 to make the models sufficiently expressive to differentiate between groups. Results are displayed in Figure 1.

As is apparent in the figures, enforcing fairness through demographic parity, True Negative Rate, and True Positive Rate is achieved by increasing performance for the disadvantaged group and reducing performance for the advantaged group (e.g., "Male" for demographic parity and True Positive Rate, and "Female" for True Negative Rate).

Similar behavior was observed with IBM AI Fairness 360. Compared to FairLearn, IBM's toolkit is less cohesive because it is comprised of a collection of different pieces of research code written by many different authors. Surprisingly, some of the code samples provided in the toolkit failed to improve fairness on their own training set. We were, however, able to obtain results for one code sample that did not suffer from this weakness: reweighting pre-processing[57] on the UCI German Credit Dataset.[58] These results are displayed in Figure 2.

Again, parity was achieved by both increasing performance for the disadvantaged group (i.e., "Female") and decreasing performance for the advantaged group (i.e., "Male"). However, despite the obvious decreases in performance for some groups in the above examples, the default reporting standards in fairML make them impossible to identify. Fairness reporting tends to reduce what should be at least two measures of group performance (e.g., selection rate or recall per group) into one (i.e., a measure of inequality such as difference in the selection rate or recall between groups). This simplification makes it impossible to determine who is harmed, and who, if anyone, is helped by the enforcement of group fairness.

[53] Agarwal et al., *supra* note 28, at 67.

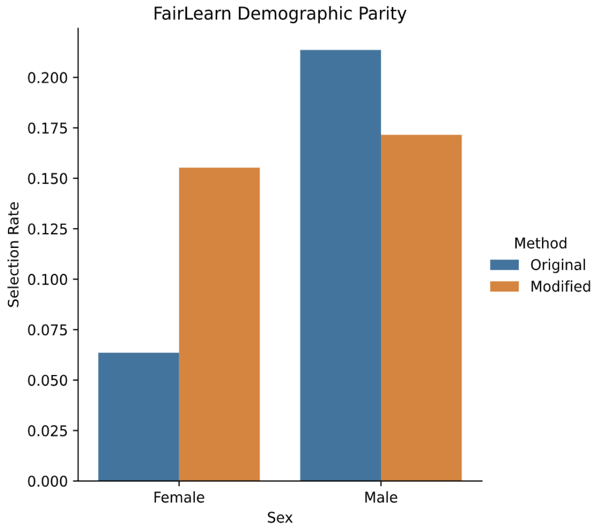
[54] *See* Fabian Pedregosa et al., *Scikit-Learn: Machine Learning in Python*, 12 J. MACH. LEARNING RSCH. 2825 (2011).

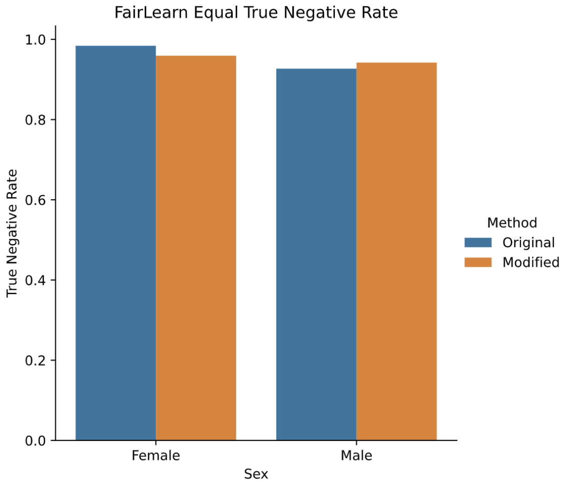
[55] Barry Becker & Ronny Kohavi , *Adult*, UCI MACHINE LEARNING REPOSITORY (2019), http://archive.ics.uci.edu/ml.

[56] Hardt et al., *supra* note 28, at 3–4.

[57] Kamiran & Calders, *supra* note 28.

[58] Dua & Graff, *supra* note 51.


FairLearn Demographic Parity
Selection Rate
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Female
Male
Sex
Method
🟦 Original
🟧 Modified


FairLearn Equal True Negative Rate
True Negative Rate
0.0
0.2
0.4
0.6
0.8
1.0
Female
Male
Sex
Method
🟦 Original
🟧 Modified

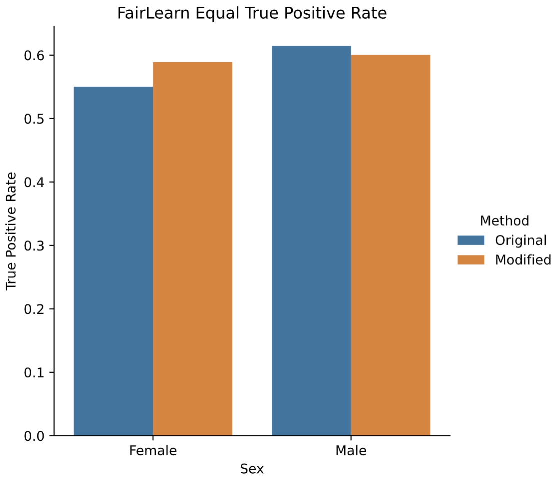


Figure 1 - Leveling down in FairLearn on the UCI Adult Dataset

However, even after identifying the drop in performance, it remains an open question whether these are genuine cases of leveling down, or cases in which the loss of performance for advantaged groups is motivated by mathematical convenience only and is not causally necessary to achieve the increase in performance for disadvantaged groups. This question cannot be conclusively answered due to the aforementioned limitations of reporting standards in the field.[59]

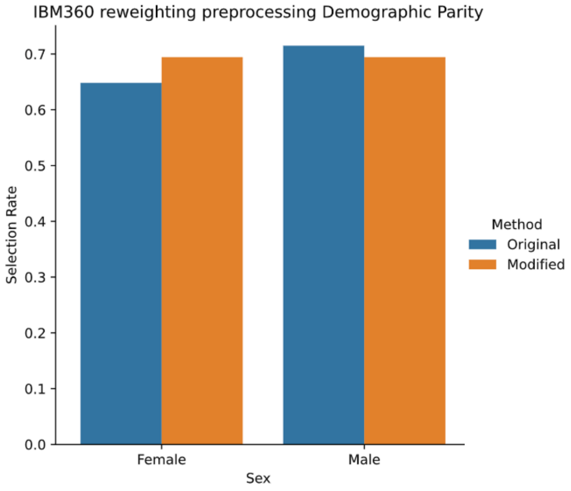


Figure 2 - Leveling down in IBM AI Fairness 360 on the UCI Adult Dataset

[59] *See supr*a Section II.B.

The field's focus on minimizing inequality while maximizing accuracy has left us without the tools needed to achieve fairness purely by leveling up, which would mitigate avoidable material harms, such as stigmatization and the loss of solidarity for both advantaged and disadvantaged groups.[60] Simply put, given how fairness is currently enforced and reported in ML, we cannot determine if harming particular groups is in fact necessary and justified, or merely the path of least resistance to achieve parity. However, in Section V we propose new leveling up tools for algorithmic fairness and show that they can reduce harms while improving performance for disadvantaged groups without disadvantaging others.

C. Leveling Down in Theory

The limitations of current reporting methods make it difficult to determine the frequency and justifiability of leveling down in fairML. An alternative approach is to determine whether the theoretical foundations of popular fairness measures consider leveling down to be a legitimate distribution mechanism. This line of inquiry cannot, of course, establish the empirical prevalence of leveling down, but can indicate whether leveling down is theoretically coherent to enforce group fairness measures.

Fairness measures are implicitly inspired by, or explicitly derived from, theories of distributive justice.[61] Distributive justice is an umbrella term describing a sub-field of political philosophy that encompasses a range of theoretical approaches to question of justice. The field addresses the "relative impact of allocations on different social groups or subgroups within the population, given existing social inequalities."[62] Justice can be conceived of in comparative or absolute terms, meaning we may be concerned simply because people are treated differently or unequally, or, alternatively, because their current treatment does not provide them with the basic goods they deserve, such as a minimum level of welfare or human rights.[63] Along these lines, theories of distributive justice specify how goods and burdens should be distributed among individuals and groups.[64] They tend to address allocation of

[60] See *infra* Section III.C.

[61] Binns, *supra* note 3, at 2; Matthias Kuppler et al., *Distributive Justice and Fairness Metrics in Automated Decision-Making: How Much Overlap Is There?*, ARXIV 17 (May 6, 2021), http://arxiv.org/abs/2105.01441 (last visited July 13, 2022).

[62] Hoda Heidari et al., *On Modeling Human Perceptions of Allocation Policies with Uncertain Outcomes*, ARXIV 4 (May 6, 2021), http://arxiv.org/abs/2103.05827 (last visited Aug 14, 2022).

[63] *See* KASPER LIPPERT-RASMUSSEN, BORN FREE AND EQUAL? A PHILOSOPHICAL INQUIRY INTO THE NATURE OF DISCRIMINATION 66 (2013).

[64] Considering group fairness in ML on the one hand and distributive justice on the other necessarily leads to some terminological confusion. As explained above we refer to advantaged and disadvantaged groups in ML according to their comparative level of performance. See *supra*

resources that are "rivalrous," meaning they are "consumed" by allocation and unavailable for further distribution, and "scarce," meaning that there may not exist an ideal amount of the resource to satisfy everyone.[65]

Group fairness measures are related to egalitarian thinking in distributive justice. Egalitarianism, one major approach within distributive justice, describes a group of theories which assign some value to equality itself.[66] Egalitarianism treats justice as a comparative concept that should be achieved through equality or reducing disparity in the distribution of a given property or resource. Group fairness measures accordingly aim to ensure "some form of statistical parity (e.g., between positive outcomes or errors) for members of different protected groups (e.g., gender or race),"[67] where groups are defined by "different values for a set of protected attributes."[68] While all group fairness measures are based on some form of statistical parity, they differ in the properties they target, such as equality of opportunities, outcomes, treatment, mistreatment, and minimal thresholds of discrimination, among others.[69] Nonetheless, all group fairness measures share an egalitarian aim of achieving parity among groups along one or more chosen properties or performance measures.

In distributive justice, equality can often only be achieved by making some groups worse off.[70] Strict approaches to egalitarianism, which value equality intrinsically and ignore other considerations such as welfare, view leveling down as justifiable. In other words, according to

Section I. These terms are also used in the distributive justice literature alongside terms such as "better off" and "worse off." The terms are related but distinct. (Dis)advantage in distributive justice can be understood in comparative or absolute terms and is measured by access to some good or benefit, and according to one's theoretical commitments may focus solely on distributions in a case at hand, or instead account for historical distributions and social factors which affect the relative value of goods for groups. In practice, these two uses of the terms overlap in practice; historically disadvantaged groups are often the groups which likewise suffer from worse performance in classification problems. *See* Obermeyer & Mullainathan, *supra* note 20. We have discussed this observation, that historical inequality is a significant consideration in questions of distributive justice and should be accounted or in fairML, at length elsewhere. *See generally* Wachter, et al., *supra* note 9. To avoid terminological confusion, we add the prefix "(historical)" whenever discussing historically (dis)advantaged groups in the context of ML problems.

[65] Kuppler et al., *supra* note 57, at 3; Derek Parfit, *Equality and Priority*, 10 RATIO 202, 202 (1997).

[66] TEMKIN, *supra* note 10, at 7.

[67] Binns, *supra* note 3, at 514.

[68] *Id.* at 515.

[69] Binns, *supra* note 3; Kuppler et al., *supra* note 57; Hu & Chen, *supra* note 24. Additionally, some approaches target the distribution of errors between groups. *See* Binns, *supra* note 3; Kuppler et al., *supra* note 57, at 11. Others are concerned with calibration or "how closely the model's estimation of the likelihood of something happening corresponds to the actual frequency of the event happening." Binns, *supra* note 3 at 515.

[70] *See infra* Section IV. *See also* Thomas Christiano & Will Braynen, *Inequality, Injustice And Levelling Down*, 21 RATIO 392 (2008); Holtug, *supra* note 10; Doran, *supra* note 10.

strict egalitarianism, it is acceptable to make a group worse off without directly benefiting others, to eliminate disparity.[71] This approach views justice strictly in comparative terms and ignores absolute entitlements. Achieving equality by making everyone worse off in absolute terms is acceptable for strict egalitarians[72] despite the fact that no individual experiences a direct benefit from equality.[73] It follows that leveling down, while intuitively problematic[74], is theoretically coherent from the view of strict egalitarianism. Methods to enforce group fairness measures based on strict egalitarianism (purposefully or otherwise) level down in some cases.

But how widespread are group fairness measures that align with strict egalitarianism? The majority of uses of the term "fairness'" in fairML are actually placeholders "for a variety of normative egalitarian considerations."[75] This is, however, a result of how fairness is conceptualized and measured rather than an explicit theoretical choice.[76] Works in fairML that propose group fairness measures tend not to link them explicitly to theories of distributive justice, or, if they do, link them in a superficial manner that fails to account for contextual factors or offer normative justification.[77] Those that do engage seriously with distributive justice have explicitly endorsed strict egalitarianism, having decided that an "adequate justification for an unequal distribution of prediction errors" is impossible for anyone to make in fairML.[78] Nonetheless, measures which conflate fairness with a strict notion of equality and equal treatment currently dominate the fairML literature.[79] Unsurprisingly, a tendency to achieve equality through leveling down

[71] Holtug, *supra* note 10, at 166.

[72] LIPPERT-RASMUSSEN, *supra* note 59, at 66.

[73] TEMKIN, *supra* note 10; Holtug, *supra* note 10 at 166; Doran, *supra* note 10; Parfit, *supra* note 10 at 6.

[74] *See supra* Section III.

[75] Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, 81 PROC. MACH. LEARNING RSCH. 1, 3 (2018).

[76] Kuppler et al., *supra* note 57; Cooper et al., *supra* note 8 at 47.

[77] Cooper et al., *supra* note 8, at 48.

[78] Kuppler et al., *supra* note 57, at 15.

[79] Kasy & Abebe, *supra* note 2, at 576; Binns, *supra* note 71, at 6–9; Cooper et al, *supra* note 8; Alejandro Noriega-Campero et al., *Active Fairness in Algorithmic Decision Making*, PROC. 2019 AAAI/ACM CONF. ON AI, ETHICS, AND SOC'Y 77, 79–80 (2019); Sanghamitra Dutta et al., *Is There a Trade-off between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing*, 119 PROC. 37TH INT'L CONF. ON MACH. LEARNING 2803, 2803 (2020); Irene Chen, Fredrik D. Johansson & David Sontag, *Why Is My Classifier Discriminatory?*, 31 ADVANCES IN NEURAL INFO. PROCESSING SYS. (2018); Michiel A. Bakker et al., *On Fairness in Budget-Constrained Decision Making*, PROC. KDD WORKSHOP ON EXPLAINABLE A.I. (2019); Hu and Chen, *supra* note 24, at 4.

has been observed for group fairness measures[80] regardless of their specific egalitarian theoretical grounding.[81]

It would seem, then, that enforcing group fairness endorses strict egalitarianism, albeit inadvertently, due to how fairness is measured rather than any purposeful theoretical choice. The dominance of a single theoretical approach owing to simple measurement constraints is highly concerning from a holistic and pragmatic view of distributive justice.

III. IS LEVELING DOWN JUSTIFIABLE?

While achieving parity in performance between groups frequently involves leveling down today, the failure to engage in serious theoretical discussion or offer normative arguments for its necessity and justifiability in specific contexts leaves its ethical, legal, and social acceptability unproven. Purposefully or otherwise, the default adoption of strict egalitarianism has led the field to a point where enforcement of group fairness creates avoidable harms for everyone involved.

Outside of fairML leveling down is not a new phenomenon; philosophy has long debated theories of distributive justice. In moral philosophy, the "Leveling Down Objection" has been advanced against a strict egalitarian approach to distributive justice.[82] Strict egalitarianism measures justice solely in terms of equality, preferring equal to unequal states. Under such a comparative approach to distributive justice, all that matters is whether the resource in question is equally distributed among individuals or groups without regard for other considerations, such as the absolute welfare of the groups in question. Reflecting this, inequality can be reduced in two ways: either by improving the situation of groups with lower performance, or by reducing the level of all groups else to be closer to the level of the worst performing group. The latter of these two options has been coined 'leveling down'.[83] According to the leveling down objection, a strict egalitarian whose conception of justice is based solely on a comparative notion of equality would favor a state in which all people are made equally worse off in terms of welfare to one in which different people have different levels of welfare.[84] Strict egalitarians

[80] Cooper et al., *supra* note 8, at 48; Hilde Weerts, Lambèr Royakkers & Mykola Pechenizkiy, Are There Exceptions to Goodhart's Law? *On the Moral Justification of Fairness-Aware Machine Learning*, ARXIV (2022), http://arxiv.org/abs/2202.08536 (last visited Jul 13, 2022).

[81] Kuppler et al., *supra* note 57, at 6.

[82] Campbell Brown, *Giving up Levelling Down*, 19 ECON. & PHIL. 111, 111–12 (2003).

[83] Holtug, *supra* note 10, at 166.

[84] Holtug, *supra* note 10, at 166. As an extreme example of the objection, Christiano and Braynen offer the example of a lifeboat with three passengers which will sink unless one passenger is thrown overboard. The principle of equality suggests that equal welfare can only be achieved by leveling down, meaning nobody is thrown overboard and all the passengers die. This egalitarian outcome would be preferable to an inegalitarian outcome in which one passenger is sacrificed so that the remaining two passengers can survive or have higher welfare.

ignore welfare entirely; parity is the only morally relevant consideration in distributive justice.

The validity of the objection and how well it "defeats" strict egalitarianism is the subject of significant philosophical debate.[85] In practice, the objection is sometimes attributed to a misreading of the principle of equality[86] or dismissed by noting that strict egalitarians are a rare breed and most "sensible egalitarians" are "pluralists,"[87] meaning they value other goods and do not measure justice solely in terms of equality.[88] At most, then, it would appear that leveling down would only be accepted as inherently good by strict egalitarians. Leveling down would be rejected by other schools of thought for whom justice is not simply a matter of equality but requires consideration of welfare, utility, priority, luck, and similar properties.[89]

This course of action clearly conflicts with a common sense understanding of justice. Christiano & Braynen, *supra* note 66, at 392–93.

[85] Brown, *supra* note 78, at 111; Christiano & Braynen, *supra* note 66, at 392; Holtug, *supra* note 10, at 166; Doran, *supra* note 10, at 61; Michael Otsuka & Alex Voorhoeve, *Equality V ersus Priority*, 65 OXFORD HANDBOOK OF DISTRIBUTIVE JUSTICE 1, 1 (Serena Olsaretti ed., 2018); *see* Larry S. Temkin, *Equality, Priority Or What?*, 19 ECON. & PHIL. 61 (2003); Richard J. Arneson, *Egalitarianism and Responsibility*, 3 J. ETHICS 225 (1999); Harry Frankfurt, *Equality as a Moral Ideal*, 98 ETHICS 21 (1987).

[86] Brown, *supra* note 78; Christiano & Braynen, *supra* note 66, at 126.

[87] Otsuka & Voorhoeve, *supra* note 81, at 20; Parfit, *supra* note 10, at 4.

[88] Brown, *supra* note 78. Brown, for example, has suggested that advocates of the leveling down objection must explain "what it means, in their view, to say that one thing is better than another in a respect" for leveling down to be a valid objection to egalitarianism. *Id.* at 113. One example of a pluralist reconstruction that combines equality and welfare to defeat the objection is the "*common good conception* of the principle of equality" which "favours states in which everyone is better off to those in which everyone is worse off." Christiano & Braynen, *supra* note 66, at 395. Pluralist egalitarians "believe that it would be better both if there was more equality, and if there was more utility." Parfit, *supra* note 10, at 4. Both equality and utility are thus given moral weight by pluralists.

[89] Pluralist egalitarian approaches, for example, recognize that there are other goods besides equality that should be considered in assessing distributive justice. These alternative goods can explain why it is better to achieve equality by making people better off than worse off and thus avoid the leveling down objection. Similarly, some egalitarians argue that strict egalitarianism is not required to achieve equality of treatment or opportunity. Binns, *supra* note 3. Rather, a maximin distribution is sufficient according to which inequalities are tolerated so long as they benefit disadvantaged groups. John E. Roemer, *Equality of Opportunity: A Progress Report*, 19 SOC. CHOICE AND WELFARE 455 (2002); JOHN RAWLS, A THEORY OF JUSTICE (Harvard Univ. Press rev. ed. 1999). Prioritarianism is concerned principally with absolute entitlements rather than comparative levels of welfare. Parfit, *supra* note 10, at 22–23. Distribution principles which maximize weighted utility across groups are preferred, with priority given to benefits to the worse off, and inequalities between groups found acceptable if they lead to greater utility. *See* Otsuka & Voorhoeve, *supra* note 81; Parfit, *supra* note 10; Parfit, *supra* note 61, at 213. Approaches to prioritarianism differ on how they measure priority and the size of benefits across groups. *See* LIPPERT-RASMUSSEN, *supra* note 59. A maximin approach, for example, focuses solely on benefits to the worst off. RAWLS; Gerald A. Cohen, *On the Currency of Egalitarian Justice*, 99 ETHICS 906 (1989). Relatedly, sufficientarianism favors distribution principles which ensure all people receive at least a minimum threshold of resources to ensure a good quality of life. Inequalities are tolerated once this minimum level is met across relevant groups. *See* Frankfurt, *supra* note 81; Jonathan Herington, *Measuring Fairness in an Unfair World*, PROC. AAAI/ACM CONF. ON AI, ETHICS, AND SOCIETY 286 (2020).

The objection helpfully draws focus to a key question for the future of fairML: under what theoretical or practical conditions, if any, can leveling down to enforce group fairness be justified?

A. The Value of Equality

Egalitarians believe equality has value. This simple statement hides significant theoretical complexity.[90] Egalitarian approaches can be distinguished according to the type of value they assign to equality. For strict egalitarians, equality has "intrinsic value," meaning it is good in itself. Conversely, inequality is bad in itself, even when it has no negative effects.[91] This intrinsic value is reflected in formal notions of equality, such as the principle of equal treatment, according to which similar individuals must be treated similarly regardless of their ethnicity, sex, gender, and other protected characteristics.[92] Prohibitions against direct discrimination or disparate treatment, for example, prohibit "less favorable" treatment on these grounds.[93]

If equality has intrinsic value, a situation is measurably improved if distributions are equalized among groups, regardless of the collateral costs of achieving parity on the groups in question. While some strict egalitarian theorists will not characterize inequality as "bad" unless certain conditions are met—for example, that it arose through no fault or choice of the individual[94]—a pure strict egalitarian would argue that the harm of inequality arises from the mere fact that some group of people are worse off than others.[95] For example, achieving the equal distribution of public housing by eliminating public housing altogether would be acceptable to a strict egalitarian because, independent of the severe impact of such a policy on tenant welfare, inequality has been removed.[96]

The intuitive problem with formal notions of equality is that equal treatment becomes a comparative measure that need not be concerned with absolute levels of welfare or benefit. For example, equal treatment

[90] Amartya Sen, *Equality of What?*, *in* THE TANNER LECTURES ON HUMAN VALUES, VOL. I (Sterling M. McMurrin ed., 1980), *reprinted in* JOHN RAWLS ET AL., LIBERTY, EQUALITY AND LAW: SELECTED TANNER LECTURES ON MORAL PHILOSOPHY 139 (Sterling M. McMurrin ed., 1987).

[91] TEMKIN, *supra* note 10, at 19–20; Parfit, *supra* note 10, at 6.

[92] Wachter et al., *supra* note 9, at 716–720; Sandra Fredman, *Substantive Equality Revisited*, 14 INT'L J. CONST. L. 712 (2016).

[93] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law a nd AI*, 41 COMPUT. L. & SEC. REV. (2021).

[94] Parfit, *supra* note 10, at 12; Cohen, *supra* note 85; TEMKIN, *supra* note 10; Richard J. Arneson, *Equality and Equal Opportunity for Welfare*, 56 PHIL. STUD. 77 (1989); THOMAS NAGEL, EQUALITY AND PARTIALITY (1991).

[95] Parfit, *supra* note 10, at 85.

[96] *Id.* at 98.

can be achieved "whether the two individuals are treated equally well or equally badly" or by "removing a benefit from the relatively privileged group."[97] However, intrinsic value alone cannot explain the intuition that it is better to achieve equality by bringing all relevant groups up to an equal level than to make them equal but worse off. For that, appeal to some instrumental value of equality is necessary.[98]

Equality can be said to have *instrumental value* derived from the good effects it produces. It is good instrumentally to achieve some other valuable goal related to justice,[99] such as universal freedom, the development of human capacities and personality, mitigation of suffering and stigmatization, or avoiding conflict or envy created by inequality.[100]

Inequality may be instrumentally bad because of the social injustice it creates.[101] There are two types of injustice to consider: (1) a comparative sense of justice, meaning we are concerned with the mere fact that people are treated differently from others, or do not receive their fair share of a resource, and (2) a non-comparative sense of justice, where injustice arises because a person is not treated as they deserve, independent of any consideration of how others are treated. With regard to leveling down, a comparative sense of justice would not view the elimination of disparity treating everyone neutrally but equally poorly as inherently unjust or problematic.[102] In contrast, a non-comparative sense would reject leveling down on the basis that people are treated poorly in absolute terms, for example being denied a vital resource.[103]

B. Performance, Utility, and Harms

Unless one is a strict egalitarian who believes in the intrinsic value of equality above all else, leveling down can only be justified by appealing to some instrumental value created by greater equality between groups that offsets the harm caused by a reduction of performance or access to a valuable good. Nonetheless, it is intuitively difficult to claim that equality has instrumental value in a case where no group experiences a direct benefit. This draws on a key intuition in moral

[97] Fredman, *supra* note 88, at 717.

[98] Otsuka & Voorhoeve, *supra* note 81, at 82.

[99] TEMKIN, *supra* note 10 at 282; Parfit, *supra* note 10, at 86.

[100] Thomas M. Scanlon, The Diversity of Objections to Inequality 1, 6 (Lindley Lecture, University of Kansas, 1996); THOMAS M. SCANLON, WHY DOES INEQUALITY MATTER? 91–92 (2018).

[101] Parfit, *supra* note 10, at 88.

[102] Discrimination, like equality, is essentially comparative—it must be possible to show that someone is better or worse off than someone else to say discrimination has occurred. *See* LIPPERT-RASMUSSEN, *supra* note 59, at 27.

[103] Parfit, *supra* note 10, at 89.

philosophy called the "person-affecting view,"[104] which says "that what makes… outcomes good and bad is how they *affect people*." [105] It follows that "because leveling down affects no-one for the better… it cannot seem to make an outcome better."[106]

The harm of leveling down can thus be linked to the inherent value of the good or benefit being distributed. Indeed, the observation that justice is not derived solely from equality but must also account for the utility of distributed goods and their impact on recipients' welfare, rights, and other interests is central to strict egalitarianism's alternatives, such as pluralist egalitarianism, prioritarianism, and welfarism. Consider, for example, the classifiers used in cancer screening discussed above.[107] A reduction in classifier accuracy for any group constitutes a welfare harm by increasing the rate of misdiagnosis for that group. In other words, performance is inherently valuable for patient health and welfare. It may, of course, be possible to justify this welfare harm if it improves performance for another worse-performing or otherwise priority group, but the existence of the harm to the welfare of the advantaged group remains constant.[108]

Furthermore, equal predictive accuracy does not guarantee equal outcomes or an equal distribution of harms and utility. For example, imagine two groups that have equal predictive accuracy but significantly different false positive and false negative rates. In healthcare this can mean that one group is more frequently misdiagnosed as healthy while the other is more frequently misdiagnosed as sick, resulting in one group being undertreated while the other is overtreated.[109] Depending on the context, either option may be preferable or problematic, as "different types of errors have different costs."[110] Overtreatment, for example, has been linked to significant harms like complications stemming from unnecessary treatments such as elective surgeries based on the results of mammography screenings.[111] By contrast, even if a diagnosis of

[104] TEMKIN, *supra* note 10, at 327; Holtug, *supra* note 10, at 167; Doran, *supra* note 10, at 65; Parfit, *supra* note 10, at 114.

[105] Holtug, *supra* note 10, at 167.

[106] *Id.*

[107] See *supra* Section I.B.

[108] The idea that performance can be inherently valuable and loss of performance harmful, while seemingly obvious, is not universally acknowledged in fairML. In a paper endorsing strict egalitarianism for fairness measures, for example, Kuppler et al. observed that "there is no obvious reason why some individuals should deserve or need a higher probability of prediction errors than others." *See* Kuppler et al., *supra* note 57, at 12. Clearly this approach ignores the real harms caused by lowering performance for advantaged groups in the search for perfect equality in error rates. *See, e.g.,* Section I.B (discussing real harms like missing more cases of cancer).

[109] Deborah Hellman, *Measuring Algorithmic Fairness*, 106 V A. L. REV. 811, 820–22 (2020).

[110] *Id.* at 829.

[111] PETER C. GØTZSCHE, MAMMOGRAPHY SCREENING: TRUTH, LIES AND CONTROVERSY (2012).

rabies comes back negative, the cost of a false negative justifies proceeding with treatment anyway.[112] Thus, the value of the distributed goods or benefits and the cost of different types of errors have a direct influence on the justifiability of leveling down in practice.[113]

Reflecting the inherent value of distributed goods, claimants in equality law cases typically do not seek a remedy that involves leveling down; rather, they seek inclusion in some benefit they are currently denied. Thus, they implicitly propose an "alternative distributive principle," which would, in their estimation, more equitably distribute the benefit. As Reaume explains, "…leveling down is rarely the remedy litigants pursue: they ask to be allowed to vote as well, not that voting be abolished, or that a pension scheme include them, not that it be repealed."[114] Doing so "would deprive everyone of something all are properly entitled to, and thus exacerbate rather than solve the problem."[115]

Leveling down is harmful because it denies access to a valuable resource to more people than is strictly necessary. Something of value is lost or removed from an advantaged group to reduce disparity. This provides no direct improvement in utility (e.g., performance, welfare) to disadvantaged groups. There is no one for whom a leveled down situation is measurably improved, and it will be actively harmful in cases where performance is inherently valuable. Its instrumental value can only be measured indirectly in terms of opportunities or benefits to disadvantaged groups in future distribution scenarios because whatever utility is lost, such as access to goods, opportunities, or other resources, is not re-distributed to disadvantaged groups. If that was the case, we would not be speaking of an instance of leveling down).

For now, we can conclude that leveling down can be rejected where it produces no instrumentally valuable direct benefits to disadvantaged groups, such as improvements in utility, welfare, or priority. It may, nonetheless, be justifiable if (1) we solely value equality above all else as strict egalitarians, or (2) can appeal to the instrumental value of its indirect effects. We have already considered the first option and will now turn to the second in examining the substantive equality harms and opportunities created by leveling down.

[112] Hellman, *supra* note 105, at 829.

[113] *Id.*; Michael Pace, *The Epistemic Value of Moral Considerations: Justification, Moral Encroachment, and James' "Will To Believe,"* 45 NOÛS 239, 257 (2011); s *ee* Brent Mittelstadt et al. , *The Ethics of Algorithms: Mapping the Debate*, 3 BIG DATA & SOC'Y 1 (2016).

[114] Denise G. Reaume, *Dignity, Equality, and Comparison*, *in* PHILOSOPHICAL FOUNDATIONS OF DISCRIMINATION LAW 7, 11 (Deborah Hellman & Sophia Moreau eds., 2013).

[115] *Id.*

C. Substantive Equality Harms

There are a range of possible harms of inequality and leveling down which cannot be strictly measured in terms of equality or solved through equal treatment. Notions of substantive equality draw attention to the subtle qualitative harms of leveling down which are not captured in simple utilitarian calculations focused on "the well-being of the advantaged group and the costs of relinquishing inequality's privileges."[116]

Focusing solely on relative disadvantages between groups ignores the "stigma, stereotyping, humiliation, and violence on grounds of gender, race, disability, sexual orientation, or other status" experienced by disadvantaged individuals.[117] Likewise, relational and identity-based harms arising from misrecognition, denigration or humiliation are not captured by utility. As Brake summarizes, the ability to "participate on equal terms in community and society more generally" is fundamentally valuable but not fully captured by strict or formal equality.[118]

Ideal solutions to inequality should fully remedy "all of the injuries, material and nonmaterial, to the persons disadvantaged by inequality," and not simply deal in redistribution of goods or resources.[119] Leveling down fails to address the social and relational harms of inequality. As explained by Fredman:

Where the injury inheres in the materially dissimilar treatment of persons otherwise similarly situated, it may be remedied by eliminating the differential treatment either by leveling down, leveling up, or setting a baseline at some point in between. Where the injuries from discrimination transcend the material consequences of differential treatment and are social or relational in nature, however, leveling down may exacerbate the injuries of discrimination and is not consistent with equality law.[120]

Restricting access, lowering performance, removing goods, or otherwise leveling down rather than leveling up expresses "unequal concern" and solidifies pre-existing social inequality, causes stigmatization and social backlash, and undermines solidarity between groups in society.[121] Historically advantaged groups may indeed prefer to surrender a benefit rather than extend access to maintain their relative privileged position and solidify "status hierarchies."[122]

[116] *See generally* Wachter et al., *supra* note 9; Fredman, *supra* note 88; Deborah L. Brake, *When Equality Leaves Everyone Worse Off: The Problem of Leveling D own in Equality Law*, 46 WM. & MARY L. REV. 513, 539 (2004).

[117] Fredman, *supra* note 88, at 730.

[118] *Id.* at 732.

[119] Brake, *supra* note 112, at 539.

[120] *Id.* at 560.

[121] *Id.* at 607.

[122] *Id.* at 578.

Social and relational harms of leveling down can also be captured in pluralist egalitarian approaches. Social and political egalitarians believe that "material and social inequalities are bad when and because they undermine individuals' ability to live as equal citizens who are willing to offer and abide by fair terms of social cooperation."[123] In this context, equality has instrumental value because its absence divides communities by engendering "morally problematic attitudes," such as "servility, envy, and a lack of self-respect among the worst off and arrogance and a jealous guarding of relative advantage among the better off."[124]

Examples of stigma and relational harms abound in equality law jurisprudence. In the U.S., *Cazares v. Barber*[125] dealt with a student who was denied entrance to the National Honor Society (NHS) on the basis of pregnancy, marital status, and living situation. The student won her case against the school district on Title IX and Fifth Amendment grounds. In response, the school district canceled the entrance ceremony and ended its participation in the NHS program. The student did not gain access to the benefit sought and caused the benefit of the program to be withdrawn from all students. Similarly, in *Heckler v. Mathews*,[126] male plaintiffs sought access to spousal benefits under the Social Security Act which were available to "wives and widows, but not husbands or widowers." According to Brake, the Court chose not to extend the benefits because "the injury in an equality claim inheres in the stigma from the discriminatory treatment and not the deprivation of the material benefit itself," meaning that the injury could be remedied by enforcing equal treatment—i.e., "denying benefits to women rather than extending them to men."[127] Both cases were remedied by enforcing equal treatment through leveling down, and thus do not reflect unequal concern for the affected groups. Nonetheless, both solutions can be characterized as causing relational and stigma-based harms to the claimant in their self-perception and relationship with others in their communities.[128]

Other classic examples of leveling down which undermine social solidarity come from the era of racial desegregation in the U.S. In *Palmer v. Thompson*,[129] rather than integrate, the city of Jackson, Mississippi responded to a court's order to desegregate by closing all its public swimming pools. Thus, rather than extend access to the

[123] Otsuka & Voorhoeve, *supra* note 85, at 65; Richard Norman, *The Social Basis of Equality*, 10 RATIO 238 (1997); Elizabeth S. Anderson, *What Is the Point of Equality?*, 109 ETHICS 287 (1997); Martin O'Neill, *What Should Egalitarians Believe?*, 36 PHIL. & PUB. AFFS. 119 (2008).

[124] Otsuka & Voorhoeve, *supra* note 81.

[125] Cazares v. Barber, No. CIV-90-0128-TUC-ACM (D. Ariz. May 31,1990), *aff'd*, 959 F.2d 753 (9th Cir. 1992).

[126] 465 U.S. 728 (1984).

[127] Brake, *supra* note 112 at 593.

[128] *Id*. at 562–563.

[129] 403 U.S. 217, 219 (1971).

disadvantaged group, the city leveled down by removing pool access for the advantaged group. (i.e., white citizens) rather than extending it to the disadvantaged group (i.e., Black citizens). The Supreme Court upheld the city's action, reasoning that there is no affirmative "right to a pool," thus, the city had no affirmative duty to operate a pool. Similarly, in *Griffin v. Cty. Sch. Bd. Of Prince Edward City*, in response to a desegregation order, the school district responded to a desegregation order by closing public schools and opening "private" schools in their place that served only white students. The Supreme Court, however, held that action constituted an equal protection violation, not because the public schools were closed, but rather because the "private" schools continued to receive support from the State and County.[130]

Leveling down to enforce equal treatment is not a uniquely American solution to inequality.[131] In *A. v. Secretary of State for the Home Department* a challenge was posed to legislation that granted "authorities the power to detain non-UK nationals indefinitely without trial if they were suspected of international terrorism." The legislation was struck down by the House of Lords on the basis that non-UK and UK nationals were "alike" in their capacity to commit terrorism and thus should be treated alike. In response, the government leveled down by extending its power for indefinite detention to both UK and non-UK nationals, ensuring equal treatment by "intruding equally on the liberty of both groups."[132]

While aligned with formal equality, the solutions in these cases can still be criticized on substantive equality grounds. Strictly speaking, leveling down in these cases treated all groups "the same in material respects."[133] But this focus solely on equal treatment misses how such solutions "express selective disdain or disregard for some persons," and reproduce or reinforce inequality through the "expressive meaning" of the judgement or corrective action.[134] One example where courts seemingly recognize the significance of expressing meaning is found in *Johnson v. California* (543 U.S. 499 2005), which dealt with the racial segregation of prisoners. There, the U.S. Supreme Court rejected the idea that racial segregation can be considered neutral or acceptable if all racial groups are equally segregated.[135] This judgement followed the spirit of the historic *Brown v. Board of Education*, where the Court found

[130] Thomas B. Nachbar, *Algorithmic Fairness, Algorithmic Discrimination*, 48 FLA. STATE UNIV. L. REV. 51, 554 n.203 (2021).

[131] It is worth noting that leveling down is not a historic relic of equality law. According to Brake, "the underlying premise–that equality law has little or nothing to say about leveling down as a response to inequality–has remained largely unchallenged" between the 1960s and early 2000s. Brake, *supra* note 112, at 520.

[132] Fredman, *supra* note 88, at 717–8.

[133] Brake, *supra* note 112, at 571.

[134] *Id.*

[135] Fredman, *supra* note 88, at 724.

that "the state's segregation… expressed a message of racial inferiority."[136]

While leveling down is intuitively problematic at a theoretical level, it has nonetheless been repeatedly upheld as a legitimate practical solution in legal frameworks that prioritize formal equality, such as US equality law.[137] The same, however, does not hold true for legal frameworks that prioritize substantive equality, such as EU non-discrimination law, where the legal history of leveling down is much more complex.[138] Nonetheless, the preceding examples show that while people seeking remedies under equality law are often left materially equal to others, they emerge substantively worse off in terms of stigmatization and social solidarity. Equal treatment through leveling down excludes disadvantaged groups from the valuable benefit sought and expresses a preference by advantaged groups for "losing the benefit rather than broadening the community of persons sharing in it."[139] Leveling down solidifies "social stratification" by ensuring the "separateness and social inequality" between advantaged and disadvantaged people remains unchanged.[140] This trend is concerning because intuitively, "a person who is harmed by discrimination and successfully prosecutes a discrimination claim should benefit from the suit and that persons should not be made worse off unnecessarily."[141] And yet, in practice, pursuing an equality law remedy instead often further entrenches inequality and social stratification rather than eliminating them.

D. Leveling Down for Social Change

The historical legal permissibility of leveling down suggests an impoverished conceptualization of equality may still lie at the basis of many equality law frameworks due to their failure to address relational and stigma-based harms of materially equal treatment.[142] However, this conclusion is not yet merited since leveling down can be used to pursue valuable second-order or indirect substantive equality benefits that are

[136] Brake, *supra* note 112, at 572–73.

[137] *See, e.g.,* Fredman, *supra* note 88; Brake, *supra* note 112. It is worth noting that leveling down may be more readily accepted under US equality law because of the predominant focus on formal equality. In jurisdictions that favor a substantive approach to equality, which focuses on equality of opportunity rather than equality of treatment, leveling down is more problematic. *See* Wachter et al., *supra* note 9, *see also* section III (on "Leveling Down Objection").

[138] *See* Wachter et al., *supra* note 9; Fredman, *supra* note 88; TARUNABH KHAITAN, A THEORY OF DISCRIMINATION LAW (2015); Sophia Moreau, *What Is Discrimination?*, 38 PHIL. & PUB. AFFS. 143 (2010); DEBORAH HELLMAN & SOPHIA MOREAU, PHIL. FOUND. DISCRIMINATION LAW (2013); Wachter et al., *supra* note 29.

[139] Brake, *supra* note 112, at 575.

[140] *Id.*

[141] *Id.* at 540.

[142] Wachter et al., *supra* note 9.

instrumentally valuable for historically disadvantaged groups. These benefits include (1) the removal of unjustified entrenched advantages or (2) the pursuit of civil action.

Considering the first, leveling down may be justifiable if used as a mechanism to eliminate an unjustified entrenched advantage. Unequal treatment in favor of the disadvantaged may be necessary in cases where equal treatment would further solidify or exacerbate an "antecedent disadvantage."[143] We refer to this as the "leveling the playing field" aim of substantive equality. It applies in cases where removing an antecedent advantage is necessary to ensure equal access to rivalrous or scarce resources or in cases where an entrenched social or institutional advantage prevents equal consideration, capabilities, or opportunities in the future.[144] Equality here can have instrumental social value by improving opportunities for relevant connected groups in a community.[145]

Consider, for example, access to employment or education. Direct material weakening of the competitiveness of advantaged groups—for example, by barring men from university degree programs—runs counter to the substantive aims of equality law. Leveling the playing field need not disrupt the advantaged group's entitlements in this way. Instead, it can remove pre-existing exclusionary standards or arbitrary barriers to access or opportunities from the decision-making or distribution process. For example, leveling the playing field could remove a college admissions requirement that applicants obtain a degree from a male-only high school.[146] Here, leveling down is justified because the injury to advantaged groups is necessary for disadvantaged groups to realize benefits.[147]

Ideally, corrective actions should not only improve equality of results, access, capabilities, or opportunities, but also address the social and institutional structures responsible for the inequality or entrenched advantages in question.[148] Take, for example, college athletics funding where male athletics programs have historically received much more funding than equivalent female programs.[149] There, because extending current funding levels to other programs would be unsustainable for

[143] Fredman, *supra* note 88, at 718.

[144] Fredman, *supra* note 88; AMARTYA SEN, INEQUALITY REEXAMINED (1995).

[145] Parfit, *supra* note 61; Parfit, *supra* note 10.

[146] Wachter et al, *supra* note 9, at 738; Brake, *supra* note 112, at 520; Kim, *supra* note 33, at 1548; Wachter et al., *supra* note 29, at 57.

[147] Another example is land ownership, which has historically been limited to certain genders or royalty; extending access would require removing this privilege from historically advantaged groups.

[148] Fredman, *supra* note 88.

[149] *See* Brake, *supra* note 112.

most colleges, "equality law should permit some leveling down to find a baseline that is not based on male privilege."[150]

Leveling the playing field is typically inappropriate in cases that deal with fundamental rights or goods, or non-rivalrous, inherently valuable goods, such as recall or accuracy in cancer screening. Extending the right to vote to women, for example, could not have been achieved by denying the right to men, and would have not been consistent with considerations of liberty.[151] Similarly, reducing recall for male patients increases undiagnosed cancer cases, and is not strictly necessary to improve recall for female patients.

Considering the second benefit, leveling down can also be used as a type of civil action to force reconsideration of problematic social and institutional norms that contribute to unequal concern. A standout example of this justification is the extension of marriage rights to same-sex couples in Benton County, Oregon. In response to a lawsuit filed to block the extension, the County leveled down by suspending all marriage licenses until the validity of the state's marriage law was resolved. As Brake explains,

> …the leveling down occurred not as resistance to the equality challenge by gays and lesbians to the state's marriage laws, but in furtherance of it… the leveling down decision was not a defensive construction of social meaning designed to reinforce a status hierarchy disparaging gay and lesbian couples. Instead, it was a tactic designed to challenge that status hierarchy and hasten the extension of marriage to gay and lesbian couples by equalizing the status of their relationships.[152]

This step was viewed as a positive civil action in support of the LGBTQ+ community's push for marriage equality at a state level. A temporary solution of leveling down across all groups by suspending all

[150] Brake, *supra* note 112 at 594–95.

[151] Reaume, *supra* note 110, at 5.

[152] Brake, *supra* note 112, at 600. Brake offers another illustrative example drawn from college athletics. Male athletes are afforded certain privileges based on a problematic notion of masculinity which Brake argues would not be appropriate to extend to female athletes. "At the prestigious level of NCAA Division I-A football, for example, it is a common practice to have the football team housed in a hotel the night before home games. The rationale typically rests on the difficulty of otherwise controlling and disciplining the players to avoid the kind of behavior that would hurt their game performance. The practice is based on a model of a male athlete who embodies a ruggedly uncontrollable masculinity and it is applied uniquely to football players. Extending such a practice to female athletes, at least on the same rationale, would make little sense. Instead, equality should require readjusting the athletic model upon which the practice is based to a gender-inclusive standard that holds all athletes responsible for their own behavior." *Id.* at 597–98.

marriage licenses ensured equal concern rather than mere equal treatment.[153]

In fairML, leveling down as civil action can force consideration of fairness in production. For example, researchers and developers can use leveling down to delay or prevent the deployment of models with unjustifiably poor performance for disadvantaged groups. By lowering performance for an advantaged group, a developer could prevent deployers from having access to a "high accuracy" or "unbiased" model, and instead force leveling up by design until the model performs acceptably well across all groups Because such a temporary reduction in performance can economically impact the deployer in cases where the advantaged group is also the largest potential or the most politically or socioeconomically powerful customer base, this leveling down empowers disadvantaged groups by making highly disparate models less commercially viable.

Of course, these possible justifications for leveling down for social change are highly contextual and depend on the specific distribution problems and policies. Here, we have outlined possible justifications and some of the considerations weighing in their favor. However, we refrain from concluding whether a particular aim justifies leveling down in all case. Applying these justifications to cases of algorithmic fairness poses additional difficulties, to which we now turn.

IV. JUSTIFYING LEVELING DOWN IN FAIRML

There are many reasons to reject leveling down in fairML, including that (1) it unnecessarily and arbitrarily harms advantaged groups in cases where performance is intrinsically valuable; (2) it demonstrates a lack of equal concern for affected groups and can undermine social solidarity and contribute to stigmatization; (3) it fails to live up to the substantive aims of equality law and fairML and squanders the opportunity afforded by interest in algorithmic fairness to substantively address longstanding social inequalities; and (4) it fails to meet the aims of many theories of distributive justice including pluralist egalitarian approaches like prioritarianism, sufficientarianism, and others. But, the question remains; when, if ever, can leveling down be justified in fairML?

If one is a strict egalitarian concerned only with equal treatment or is enforcing group fairness to satisfy a social policy or law that requires strict egalitarianism, leveling down in fairML can be theoretically justified. Beyond these straightforward situations, the arguments discussed above which could justify leveling down in practice are a poor fit for fairML. Possible justifications relate to contextual factors such as

[153] Brake, *supra* note 112, at 561; RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 272–73 (2013).

available resources, current distributions, historical inequalities, and their impacts, and aim to achieve goals such as civil action aimed at removing or forcing a society to question an unjustified pre-existing advantage. These justifications turn on the realization of an underlying goal that is independently justified within a given legal, ethical, political, or social framework (e.g., questioning heteronormativity of marriage or removing entrenched advantages). In such cases, leveling down can be an imperfect means to realize some agreed upon end of greater value, because the elimination of utility for the sake of parity has clear instrumental value.[154]

The problem is that enforcement methods and practices in fairML currently do not engage with possible justifications and criticisms of the distribution principles they produce.[155] Fairness is treated as a standardized, solvable mathematical problem. Consequently, justifying how a measure is satisfied in practice, linking it to some underlying equality goal, and exploring whether a less equal but less harmful path would be preferable are rarely part of enforcing fairness.[156] Debates in distributive justice recognize that "different people may value the same outcome or set of harms and benefits differently."[157] This observation is not reflected in the current tendency for fairML to assume "a uniform valuation of decision outcomes across different populations"[158] and use cases. Such valuation reduces a highly complex, value-laden debate and set of theories and decisions to an oversimplified, homogenous set of assumptions.

The same type of error (e.g., false positives, false negatives) can cause substantially different types of harm depending on the use case. Take facial recognition as an example. If facial recognition is used by police to identify people with outstanding warrants in crowds, the harm of a false positive is an unjustified arrest. In contrast, if facial recognition is used to track perceived compliance with visa requirements,[159] the harm is perceived non-compliance with a monitoring regime that could have significant legal ramifications like deportation. The harms of false positives and negatives likewise vary when facial recognition is used for loan decisions or job interviews. It is essential to consider the specific types and severity of harms actually suffered by affected populations if the enforcement of fairness in ML is to resemble comparable legal decisions (where, as we have seen, leveling down can be justified).

[154] Strict egalitarians may disagree with this assessment, citing the inherent value of parity, but this view is controversial in law and philosophy. *See* Section III.

[155] Kuppler et al., *supra* note 57.

[156] *See* Kasy & Abebe, *supra* note 2; Kuppler et al., *supra* note 57.

[157] Binns, *supra* note 71, at 6.

[158] *Id.*

[159] Nicola Kelly, *Facial Recognition Smartwatches to Be Used to Monitor Foreign Offenders in UK*, the GUARDIAN (Aug. 5, 2022), https://www.theguardian.com/politics/2022/aug/05/facial-recognition-smartwatches-to-be-used-to-monitor-foreign-offenders-in-uk.

Researchers, developers, and deployers of "fair" ML systems, however, do not currently engage seriously with such questions at a local level. Leveling down is not viewed as something that requires justification. At most, one need only justify the choice of fairness measure; the steps taken to satisfy it in practice are normatively irrelevant.[160] At best, tenuous connections are drawn between fairness measures and complementary political or ethical theories.[161] Works that merely link methods and measures to complementary theories of equality suggest that researchers using those methods and measures believe in those theories, or have explicitly chosen them, and have critically thought about the theory of equality their models should promote.[162] This cannot be taken for granted. Rather, the vast majority of cases involving leveling down are unintentional and invisible, resulting from convenience and the limited ways performance and parity are currently measured,[163] not theoretical conviction or justification.

It is likewise unclear whether substantive equality goals can be achieved directly by enforcing fairness measures on ML models. Intentional leveling down can draw attention to a problematic performance gap and prompt civil action. Leveling down, however, does not produce this effect by itself because rigid enforcement of fairness measures does not allow for external corrective action to reduce harm—measures must be solvable with the data at hand.[164] For example, where resources are limited, a bigger "piece of the pie" requires a smaller piece be given to someone else. FairML, however, typically cannot make this sort of trade-off explicitly because models have neither a picture of "how big the pie is" nor awareness of resource scarcity.[165] Nonetheless, actions which are not directly quantifiable or within the control of the

[160] *See generally,* Wachter et al., *supra* note 29; Wachter et al., *supra* note 9.

[161] *See* Kuppler et al., *supra* note 57; Binns, *supra* note 3. In that sense they may inform model development or the scope of a research study but are not offered as a justification for enforcing group fairness in specific cases.

[162] Binns, *supra* note 3. We have made a similar observation in prior work introducing the notion of bias preservation, where we argued that researchers, developers, and deployers of ML systems need to explicitly choose the biases their models should exhibit. *See* Wachter et al., *supra* note 9. Our argument here is complementary but distinct; we argue that people working in fair ML need to be more explicit and reflective about the underlying goals their choice of fairness measures and methods supports.

[163] Kuppler et al., *supra* note 57.

[164] Binns, *supra* note 71.

[165] Enforcing fairness typically involves balancing output rates (e.g., acceptance rates, sufficiently high recall for cancer detection) between groups. For specific instances of leveling down in fair ML to be justified under something like the 'leveling the playing field' argument, models would need to be distributing a known limited quantity of a set of outputs. *See* Section D. In practice, this is not how classifiers operate. This observation does not, however, preclude justification of leveling down at a general level. Decision-makers may, for example, choose to lower performance for specific historically advantaged groups for all classifiers used in a given sector or for classifiers considering specific historically disadvantaged groups in order to level the playing field.

modeler, such as collecting more data on equality-relevant features (e.g., socioeconomic status, prior opportunities) or increasing available resources (e.g., cancer screening) in the production environment, are typically not considered but could be viable means to avoid leveling down in practice.[166]

To bring fairness in machine learning out of a testing and research environment and into systems that make important real-world decisions,[167] theoretically richer and contextually sensitive work cannot be the only answer. Instead of simply leveling down out of convenience to "solve" fairness and arbitrarily harming people in the process, fairML should shift to a harms-based framing. Such a framing forces case-based discussion of the justifiability of leveling down while also opening up an alternative set of solutions that "level up" systems by design.

V. LEVELING UP BY DESIGN WITH MINIMUM RATE CONSTRAINTS

Leveling up can be understood as a new type of constraint for fairness that provides an alternative to strict egalitarianism achieved through leveling down. If we believe that particular groups are harmed by decision or recall rates that are too low, we can simply increase performance to the required level. Instead of requiring that harms be equalized between groups, leveling up requires harms to be reduced to, at most, a given level per group. For example, if we believe that people are being harmed by low selection rates, precision, or recall, instead of enforcing that these properties be equalized across groups, we can instead require that every group has, at least, a minimal selection rate, precision, or recall. We refer to this type of minimum acceptable threshold for harms visited upon groups in the pursuit of fairness as a "minimum rate constraint" (MRC).

Unlike existing methods that enforce strict equality, there is no direct gain from decreasing the rate for any group when fairness is defined in terms of minimum rate constraints. Rather, the focus of fairness is on leveling all groups up to a minimally acceptable performance threshold.

[166] Binns, *supra* note 71; Cooper et al., *supra* note 8. Collection of data on equality-relevant features is not the same as collecting more representative data to combat bias against data impoverished groups, which is a common approach in the field and can avoid the need for leveling down. This is sometimes called 'active fairness'. *Id.* This approach helps mitigate biases in the existing data affecting disadvantaged groups without directly impacting advantaged group performance or outcomes. *See, e.g.,* Obermeyer & Mullainathan, *supra* note 20; Noriega-Campero et al., *supra* note 75; Dutta et al., *supra* note 75; Chen et al, *supra* note 75; Bakker et al., *supra* note 75; Lucas Dixon et al., *Measuring and Mitigating Unintended Bias in Text Classification*, PROC. 2018 AAAI/ACM CONF. ON AI, ETHICS, AND SOC'Y 67 (2018); Ruchir Puri, *Mitigating Bias in AI Models*, IBM RSCH. BLOG (2018); Heidi Ledford, *Millions of Black People Affected by Racial Bias in Health-Care Algorithms*, 574 NATURE 608 (2019).

[167] Bakalar et al., *supra* note 27.

Performance reductions are only tolerated if they are causally necessary to improve the situation of another group.

We use post-processing to show how leveling up can be achieved through MRCs in practice. The family of post-processing methods we consider[168] tune a separate offset for each group, which in turn alters the proportion of individuals in each group that receive a positive decision. By altering these thresholds, it is possible to enforce (either approximately or exactly) a wide range of fairness measures. Through changes to these thresholds we obtain a set of classifiers with varying accuracy and fairness. After discarding all bad classifiers that are both less accurate and less fair than at least one other classifier in the set, we obtain a *Pareto Frontier* consisting of the classifiers with the best possible fairness and accuracy trade-offs.

What follows are two examples where we evaluate fairness against accuracy on the UCI Adult Dataset to show how leveling down results from enforcing fairness measures as currently conceived and how it can be avoided by shifting focus from parity to minimum thresholds—what we refer to as "minimum rate constraints."[169] First, we use the standard equality notion of demographic parity as our fairness measure and show how it induces leveling down. We then compare this result to a second option where we trade off the minimum selection rate for each group against accuracy. This approach induces demographic parity without leveling down. These demonstrations were prepared using the AutoGluon-Fair fairness toolkit.[170]

A. Example 1: Demographic Parity

Figure 3 shows a Pareto Frontier for accuracy and demographic parity on the Adult Dataset. Following it are the results from the training set where demographic parity (Example 1) and true negative rate (Example 2) are enforced. Transferring them to the unseen test data introduces noise which would make the results less clear.

[168] Corbett-Davies et al., *supra* note 13.

[169] *See supra* Section II.B.

[170] Chris Russell, Weisu Yin & Nick Erickson, *Auto Gluon-Fair*, https://github.com/autogluon/autogluon-fair.

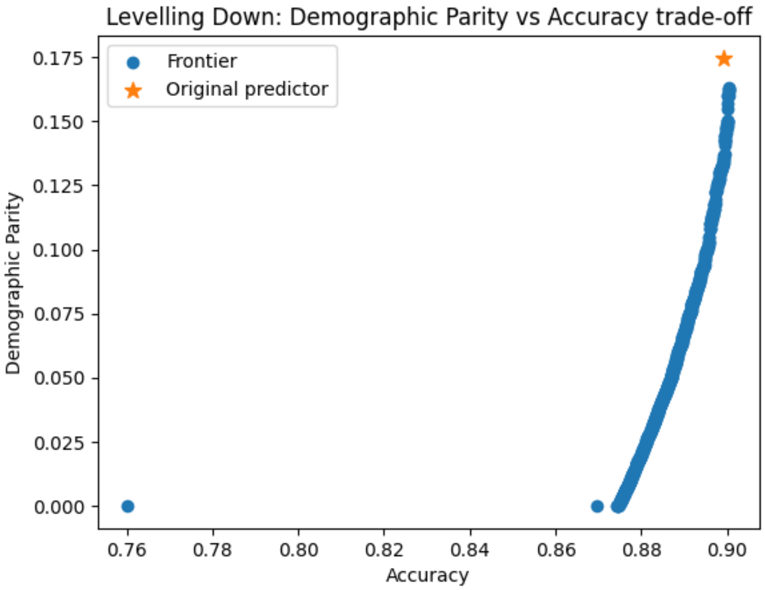


Figure 3 - Tradeoff of demographic parity vs. accuracy when enforcing demographic parity in Example 1

The dot on the far left represents a constant classifier that is perfectly fair. Figure 4 presents the computed selection rate per group for every classifier on the frontier. As expected, enforcing demographic parity exhibited leveling down evidenced by the selection rate for the advantaged group continually decreasing.

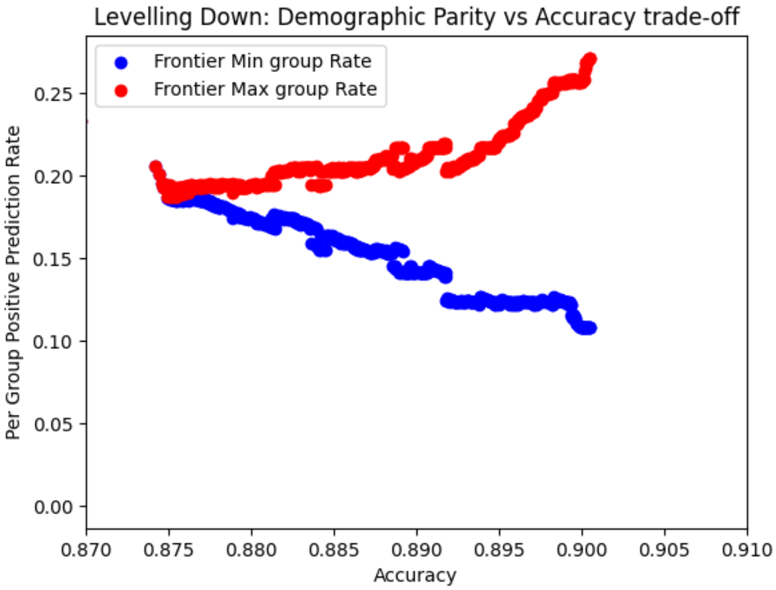


Figure 4 - Tradeoff of positive prediction rate vs. accuracy when enforcing demographic parity in Example 1

By contrast, when we instead enforced the requirement that the minimal selection rate for any group must be above a particular threshold, we observed the very different behavior visible in Figure 5.

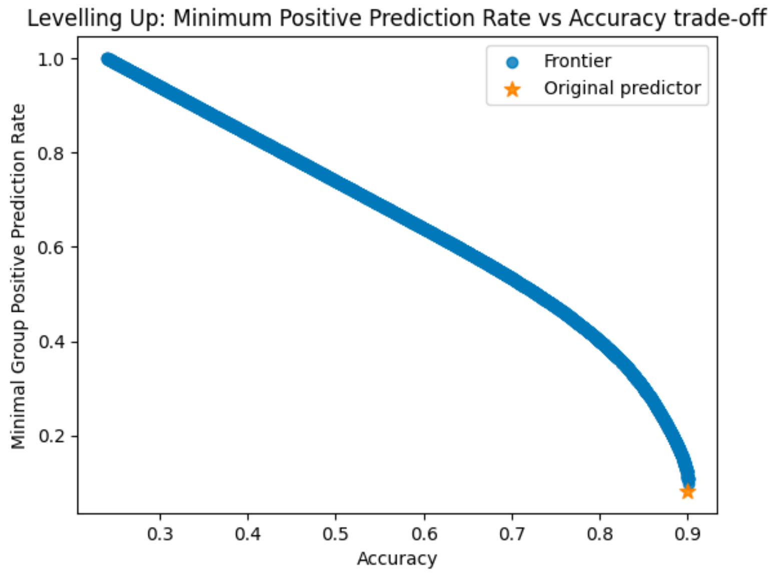


Figure 5 - Tradeoff of minimum group positive prediction rate vs. accuracy when enforcing minimum positive prediction rate in Example 1

As expected, with the dataset being more than 75% negatively labeled, large drops in accuracy were required for the positive decision rate to approach 1. Figure 6 shows the positive prediction rate for each group. Unlike enforcing egalitarian group fairness constraints, leveling down does not occur. Instead, the decision rate of the disadvantaged group steadily increases until it reaches parity with the advantaged group, followed by the decision rate for both groups increasing together.[171]

[171] To read this graph, start in the bottom right corner that shows an initial solution of high accuracy and low decision rate, and follow it to the top left corner of lower accuracy and higher decision rate.

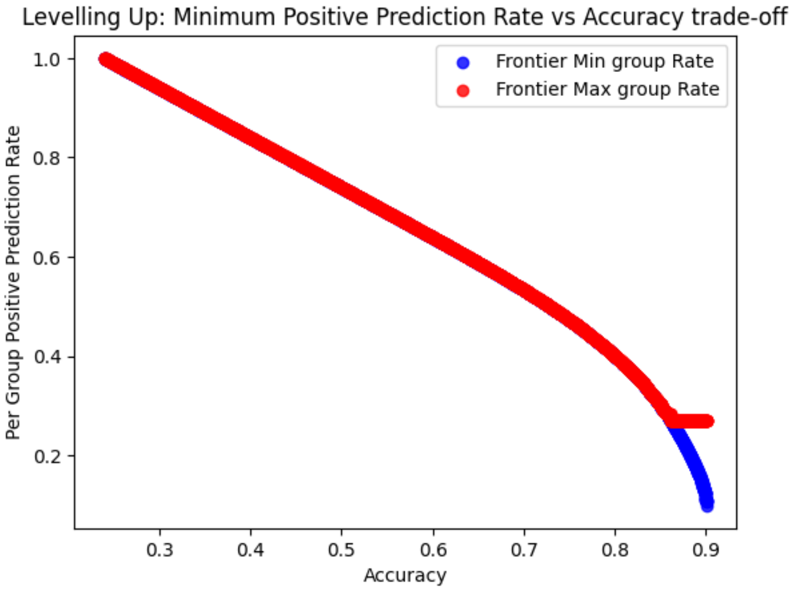


Figure 6 - Tradeoff of per group positive prediction rate vs. accuracy when enforcing minimal positive prediction rate in Example 1

As can be seen in Figure 7 the plots of the demographic parity for the frontier below, demographic parity is decreased without leveling down until parity is reached. and then it is consistently near zero, as the selection rate increases for all groups.

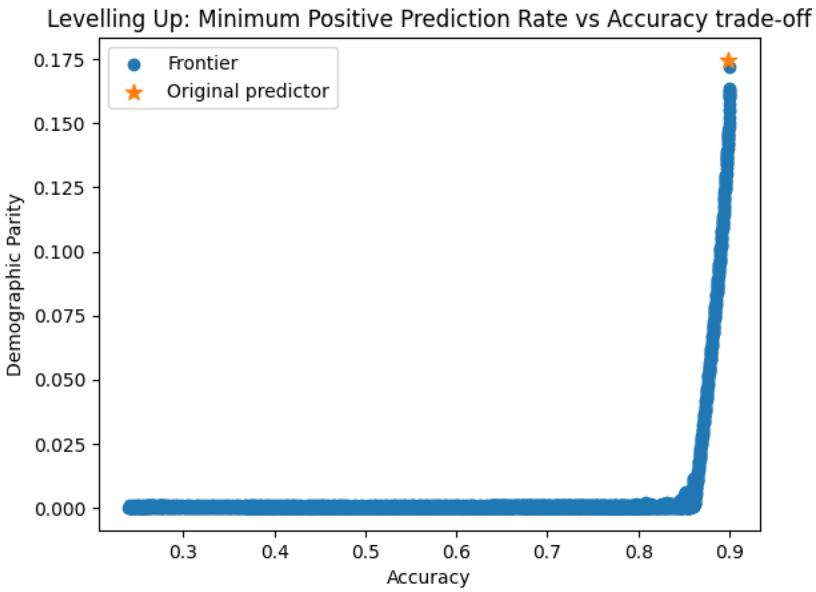


Figure 7 - Tradeoff of demographic parity vs. accuracy when enforcing minimum positive prediction rate in Example 1

B. Example 2: Difference in True Negative Rate

The same behavior can be observed for other choices of equality metric. Figure 8 below shows the same behavior for difference in true negative rate (or what is also called "false positive error rate balance" or "predictive equality").[172] Results for equal opportunity (i.e., difference in true positive rate) have similar behavior, but owing to the small proportion of datapoints with a positive label, the frontier has fewer than 10 points, making the plot much less clear for our purposes.

[172] Verma & Rubin, *supra* note 13, at 4.

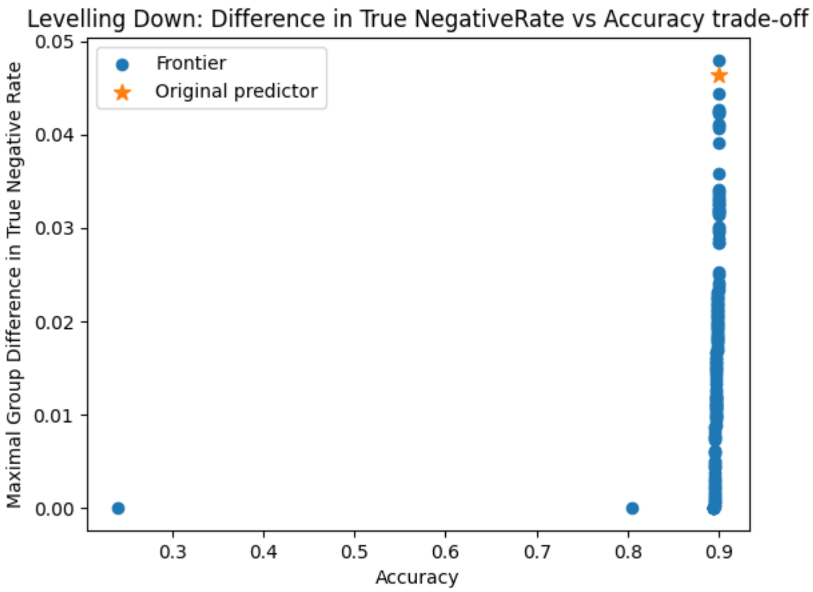


Figure 8 - Tradeoff of true negative rate vs. accuracy when minimizing difference in true negative rate in Example 2

A close-up plot showing leveling down in the vertical component of the above frontier is shown in Figure 9 below. While the curve is noisier than that for demographic parity in Figure 7, a general trend of the true negative rate decreasing for the advantaged group is nonetheless apparent.

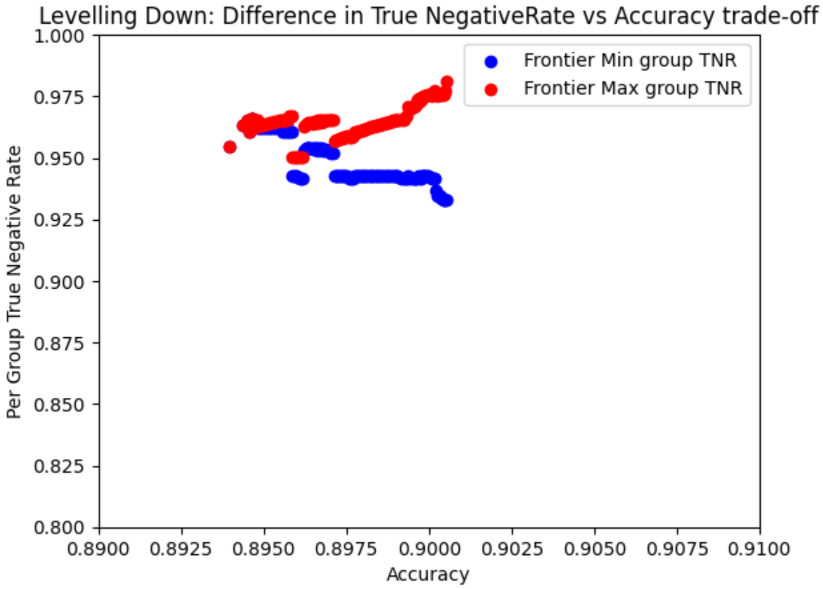


Figure 9 – Tradeoff of per group true negative rate vs. accuracy when minimizing difference in true negative rate in Example 2

If instead we compute the frontier for minimum true negative rate against accuracy, we see the frontier exhibited in Figure 10.

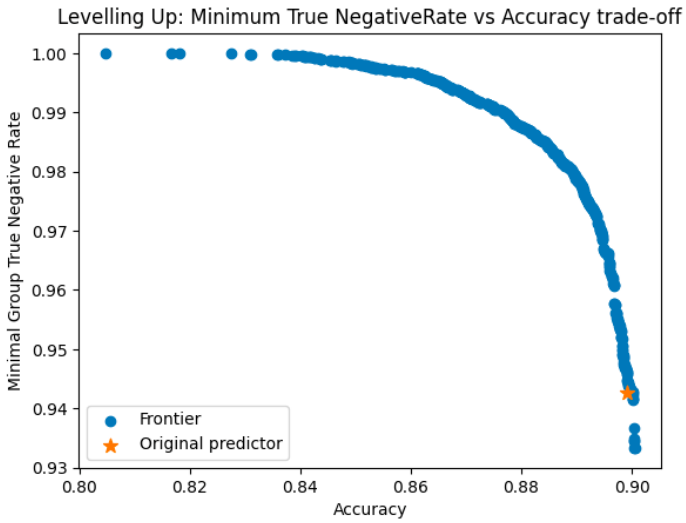


Figure 10 - Tradeoff of minimum true negative rate vs. accuracy when maximizing true negative rate in Example 2

Plotting the per group response shows the following leveling up behavior in the two plots comprising Figure 11.

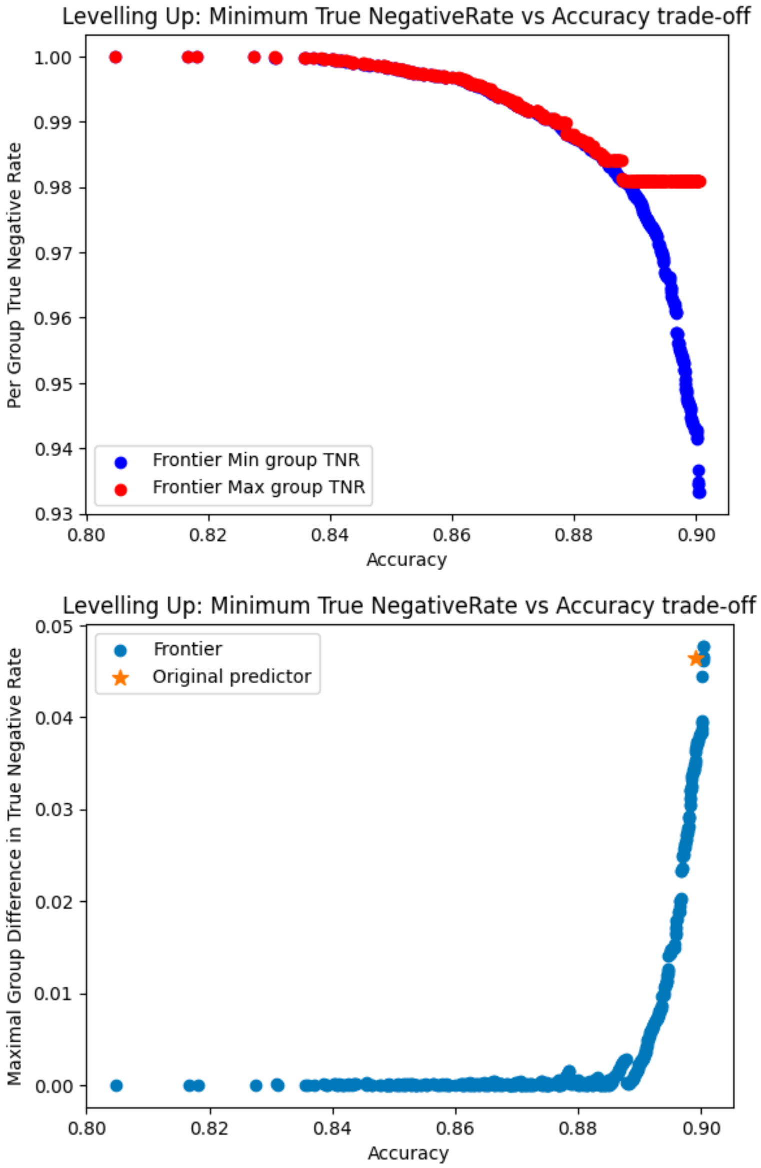


Figure 11 - Tradeoff of per group true negative rate vs. accuracy when maximizing true negative rate in Example 2

This exhibits the expected behavior of decreasing the difference in true negative rate until the disadvantaged group has a near equivalent true negative rate to the advantaged group, after which the difference remains close to zero.

Directly comparing the behavior of standard equality constraints with the MRCs (i.e., leveling up) as plotted above reveals a few insights. First, for every possible inequality value, there is a corresponding choice of minimum rate threshold that will make the inequality smaller than, or of the same size, as this value. However, because these MRCs cannot be satisfied by leveling down, a solution found by leveling up will typically have lower accuracy for a given equality level than solutions based on standard equality constraints. Nonetheless, MRCs can be informative with much smaller values than are needed to enforce standard equality constraints.

Taking this comparison further, Figure 12 compares leveling up with standard (or accuracy-maximizing) demographic parity on the Adult Dataset. The blue bars show the overall positive decision rate for an unconstrained classifier that makes positive decisions at a substantially lower rate for women than men. By enforcing demographic parity, standard fairness methods reduce harms to women at the group level at the cost of increasing harms to men. By contrast, the green bars show an example of achieving demographic parity by leveling up the positive decision rate for women until parity is achieved without altering decision rates for men. Finally, the red bars show an example of partial leveling up (as discussed above) where the decision rate for women is improved to the same level enforced by standard demographic parity without also decreasing the decision rate or accuracy for men. It is interesting to note that only standard demographic parity results in a drop of accuracy for men, and that the results for women show that it is possible to substantially increase the positive decision rate for women while maintaining a level of accuracy above or comparable to men.

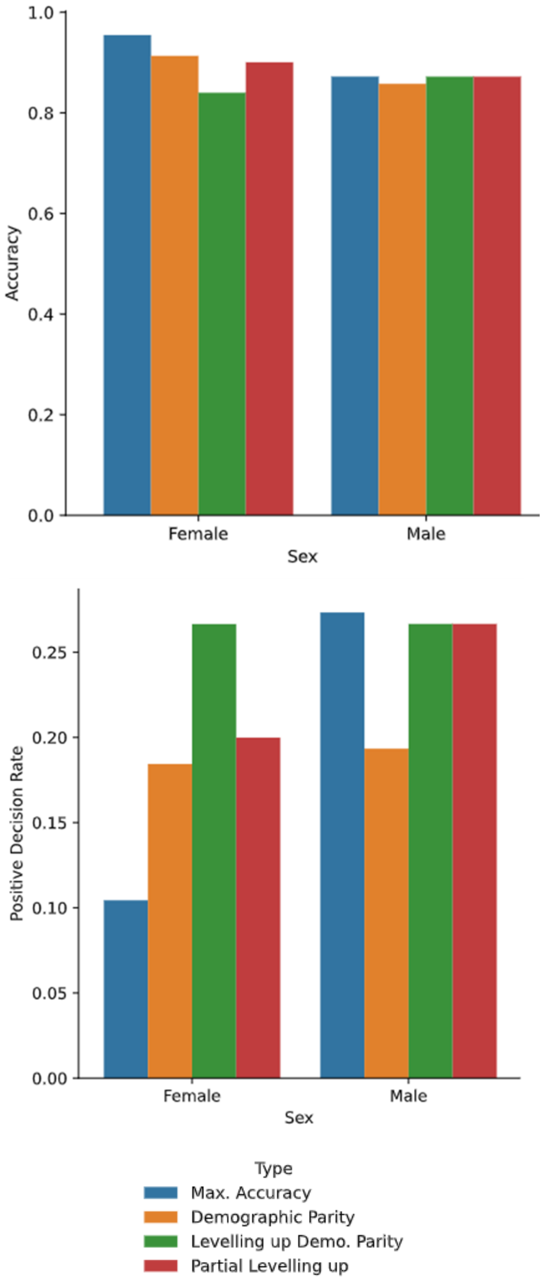


Figure 12 - A comparison positive decision rates and accuracy when enforcing demographic parity either as an equality constraint or through leveling up

The green and red bars effectively represent a choice of different MRCs for positive decision rates. Maximising this rate, represented by the green bar, has a higher cost in accuracy than partial leveling up, represented by the red bar), but it achieves parity between men and women without leveling down. In contrast, partial leveling up retains more accuracy (even compared to demographic parity) but with a lower minimum positive decision rate threshold. Which of these MRCs is preferable or justifiable in a given use case or dataset is not immediately evident. However, this is precisely the point—this type of normative decision cannot and should not be made solely from the perspective of what is easiest or most mathematically convenient. The decision should not be divorced from normative and context-specific considerations.

Leveling up avoids the "pathway of least resistance" by changing the goal of fairness from achieving strict parity between groups to ensuring a minimum "fair" performance threshold is met; it helps guarantee that reductions in performance will only be used when they are causally necessary to improve performance for another group; and it reframes fairness in machine learning in terms of harms rather than strict parity. By doing so, it necessitates local discussion at the level of specific use cases to determine normatively acceptable levels of harm and to identify additional steps to be taken or resources to be allocated to ameliorate those harms in practice. Bridging this gap between current practices and normative ideals is precisely where the real work of achieving substantive equality through fairML should begin.

CONCLUSION

Leveling down is a symptom of the choice to measure fairness solely in terms of disparity between groups and assume uniform value for benefits and harms. It is a symptom of ignoring welfare, priority, and other goods as well as substantive harms like stigmatization, unequal concern, and loss of solidarity, all of which are central to questions of equality in the real world. Our examination of group fairness enforcement methods, philosophical theories, and equality law jurisprudence shows that leveling down is not a satisfactory solution to distributive justice problems in AI and ML. We call on researchers, developers, and deployers to engage seriously with the messy socioeconomic, legal, and philosophical details of the distributive justice problems to which fairML measures and methods are applied.

Substantively improving classifier performance, when compared to leveling down, can be difficult as well as time- and resource-consuming. It may require new data, and it is not always possible for well-designed systems. Leveling down is nonetheless not the inevitable fate of enforcing fairness; rather, it is the result of taking the easier path out of mathematical convenience, and not any overarching societal, legal, or ethical reasons. Fairness cannot continue to be treated as a simple mathematical problem.

Moving forward, we see three possible pathways for fairML:

1. We can continue to deploy biased systems that benefit only one privileged segment of the population while harming others.
2. We can continue to define fairness in formalistic mathematical terms and deploy AI and ML systems that perform worse for all groups and actively harmful for some groups.
3. We can take action and achieve fairness through "leveling up," meaning we design systems to purposefully generate more false positives for (historically) disadvantaged groups and dedicate the necessary additional resources to offset the errors (for example, by increasing the frequency of cancer screenings).[173]

Throughout this paper, we have outlined many reasons to reject leveling down in fairML.[174] It shows unequal concern for disadvantaged groups, undermines social solidarity, and increases stigmatization. It causes unnecessary harm for advantaged groups in cases where a false negative can be incredibly costly in terms of health, welfare, and opportunities. But more than anything, it represents a missed opportunity to use AI and ML systems to live up to the substantive aims of equality and force reconsideration of deeply embedded inequality in the status quo.

In our view, only fairness achieved through leveling up is morally, ethically, and legally satisfactory. Leveling up is a more complex challenge: it needs to be paired with active steps to root out the real life causes of biases in AI systems. Technical solutions are often only a plaster to deal with a broken system. But to fix the system, technical solutions need to be coupled with actions to achieve substantive equality. Improving access to healthcare, curating more diverse data sets, determining the true subjective value of benefits and harms to affected populations, and developing tools that are designed with disenfranchised communities in mind are some of the steps that must be taken.

This is the challenge for the future of fairness in AI: to create systems that are substantively fair through leveling up, not only procedurally fair through leveling down. This is a much more complex challenge than simply tweaking a system to make two numbers equal between groups. It may require not only significant technological and

[173] For example, modern definitions of algorithmic fairness, such as equal opportunity, can also be satisfied by "leveling up," or increasing the rate of cancer diagnosis until the recall is the same for every demographic group.

[174] *Supr*a Sections III and IV.

methodological innovation, including re-designing AI systems from the ground up, but also substantial social changes in areas such as healthcare access and expenditures.

Difficult though it may be, this refocus on fairness through leveling up is essential. AI systems make life-changing decisions. Choices about how they should be fair, and to whom, are too important to reduce solely to a solvable mathematical problem. This is the status quo which has resulted in fairness methods that achieve equality through leveling down.

This is not enough. Existing tools are treated as a "solution" to algorithmic fairness, but thus far they do not deliver on their promise. Their morally murky effects pose a barrier to real solutions to these problems. We have created methods that are mathematically fair, but do not benefit the worst off. What we need are systems that are fair through leveling up, that help historically disadvantaged groups without arbitrarily harming others. This is the challenge fairML must now solve. We need AI that is substantively, not just mathematically, fair.

APPENDIX 1: HARMS AND REMEDIES FOR GROUP FAIRNESS MEASURES

Fairness measures	**Justified use**	**Example**	**Direct harm to individuals**	**Direct remedy**
(Conditional) Demographic Parity (or statistical parity)	Situations where historic data is expected to be prejudicial, and there is no agreed upon ground-truth.	Hiring, offering loans, access to education, representation in the media.	Lack of selection.	Increase or decrease selection rate.
Equal Opportunity (or False negative error rate balance)	Situations where there is agreed up on ground-truth and the overwhelming harm comes from false negatives.	Cancer or other serious illness screening.	Failure to identify positive cases.	Increase the recall.
*Predictive Parity**	Situations where there is agreed up on ground-truth and the overwhelming harm comes from false positives.	Misidentification as a known person of interest to the police.	Failure to identify negative cases.	Increase the precision.
*False positive error rate balance**	Situations where there is agreed up on ground-truth and the overwhelming harm comes from false positives.	Misidentification as a known person of interest to the police.	Failure to identify negative cases.	Increase the true negative rate.
*Equalized odds**	Combination of Equal Opportunity and False positive rate.	Treatment of illness by performing risky surgery.	Harms exist for failure to correctly identify positive and negative cases but they cannot be directly compared.	Increase recall and true positive rate simultaneously (may not be possible).
*Conditional use accuracy equality**	Combination of predictive parity and false positive error rate balance.	Treatment of illness by performing risky surgery.	Harms exist for failure to correctly identify positive and negative cases but they cannot be directly compared.	Increase precision and specificity simultaneously (may not be possible).

Overall accuracy equality	Situations where there is agreed up on ground-truth and the harm of misclassification is the same regardless of how people are situated.	Offering someone left- or right-handed scissors.	Harms exist for failure to correctly identify positive and negative cases and they are the same in both cases.	Increase overall accuracy simultaneously (may not be possible).
Treatment Equality	Unclear	-	--	-

* Predictive Parity and False Positive error rate balance both treat false negatives as a harm, but they normalize the harm differently. Predictive Parity is analogous to measuring the proportion of people in jail that are innocent, while False Positive Error Rate is analogous to measuring the proportion of innocent people that are in jail. While both measures relate to the number of innocent people in jail, whether this is recorded as a proportion of the people in jail, or of the total number of innocent people can drastically change what is seen as a significant harm. A similar relationship occurs between Equalized odds and Conditional Use Accuracy. Both are concerned with the same harms, but the way they are normalized varies. All fairness measures based on the classification scheme of Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, 2018 IEEE/ACM Int'l Workshop on Software Fairness (FairWare) 1 (2018).