CONTEXTUAL FAIRNESS: A LEGAL AND POLICY ANALYSIS OF ALGORITHMIC FAIRNESS

Doaa Abu-Elyounes†

Abstract

To date, all stakeholders are working intensively on policy design for artificial intelligence. All initiatives center around the requirement that AI algorithms should be fair. But what exactly does it mean? And how can algorithmic fairness be translated to legal and policy terms? These are the main questions that this paper aims to explore. Each discipline approaches those questions differently. While computer scientists may favor one notion of fairness over others across the board, this paper argues in favor of a case-by-case analysis and application of the relevant fairness notion. The paper discusses the legal limitations of the computer science (CS) notions of fairness and suggests a typology of matching each CS notions to its corresponding legal mechanism. The paper concludes that fairness is contextual. The fact that each notion, or group of notions, correspond with a different legal mechanism, makes them suitable for a certain policy domain more than others. Thus, throughout the paper, examples for possible applicability of the CS notions to some policy domains will be introduced. In addition, the paper will highlight for both developers and policymakers the practical steps that need to be taken in order to better address algorithmic fairness.

In some instances, notions of fairness that seem, on their face, unproductive from a technical perspective, could in fact be quite helpful from a legal perspective. In other instances, desirable notions in the eye of computer scientists could be challenging to implement in the legal regime, due to the need to determine complex moral and legal questions. Thus, as the article

† **Author -** Harvard Law School; Berkman Klein Center for Internet and Society at Harvard University (email: dabuelyounes@sjd.law.harvard.edu). I would like to express my gratitude to my advisor, Professor Yochai Benkler, and to Professor Cynthia Dwork, Professor Martha Minow and Professor Niva Elkin-Koren for an extremely helpful guidance throughout the writing process. I would also like to thank Christina Ilvento, Yonadav Shavit, Boriana Gjura, Meena Jagadeesan, Joshua Simons and Zhun Deng for helpful comments on the technical CS aspects of the paper, as well as the participants of Ethical Tech working group at Berkman Klein Center for their useful advice. Earlier drafts of this work were presented in the Cambridge International Law Conference and the AI and Law Conference in Montreal and I am grateful for helpful comments provided by the participants. This work was supported by a grant from the Center for Cyber Law and Policy at the University of Haifa, Israel.

emphasizes, a one-size-fits-all solution is not applicable for algorithmic fairness. Rather, an approach that demonstrates a deep understanding of the specific context that a certain algorithm is operating in can guarantee a fairer outcome.

TABLE OF CONTENTS

I. INTRODUCTION

In the summer of 2016, a story about a racist algorithm named COMPAS used in the criminal justice system blew up in the media. COMPAS is a proprietary actuarial risk and needs assessment tool, developed by Northpointe (now owned by Equivant), and is being used by criminal justice agencies in many jurisdictions in order to determine defendants' risk to recidivate.[1] Due to its proprietary nature, the inner workings of COMPAS, as well as the way the score is calculated, are not known to the public or to defendants.[2] The media

[1] EQUIVANT, PRACTITIONER'S GUIDE TO COMPAS CORE 1, 1 (2019), http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf.

[2] Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

outlet ProPublica examined the fairness of COMPAS's classification and if it is prone to racial bias by studying whether defendants who were released from jail actually recidivated after two years, as COMPAS anticipated.[3] ProPublica concluded that COMPAS is racially biased because it falsely labeled black defendants as future criminals twice as much as it did so for white defendants. While among black defendants, 42% of those who were released from jail and did not commit any future crimes were wrongly labeled high-risk, among white defendants, the algorithm made the same mistake in only 22% of cases.[4] Northpointe dissented from the findings, and published their own investigation showing how COMPAS is equally fair to black and white defendants.[5] The rebuttal attracted the attention of many other news organization and researchers that published conflicting results.[6] While some academics supported ProPublica's finding and became reluctant to the use of algorithms in the criminal justice system,[7] others attributed the gap that ProPublica found to external factors such as the different base rate among black and white defendants.[8]

To date, two key points are widely known: (1) the rebuttal between ProPublica and Northpointe centers on a disagreement about the appropriate definition of fairness that should be used in the algorithm. While for ProPublica fairness meant that the algorithm should make the same type of error equally for black and white defendants, for Northpointe the algorithm was fair since in each race category the same percentage of black and white defendants recidivated;[9] and (2) both from a policy and technical perspective, satisfying multiple notions of fairness simultaneously is mutually incompatible.[10]

[3] *Id.*

[4] Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually not that Clear.*, WASH. POST (Oct. 17, 2016, 4:00 AM), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.ef319d030999.

[5] WILLIAM DIETERICH ET AL., COMPAS RISKS SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY, NORTHPOINTE INC. RES. DEP'T, 1, 2 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[6] *See, e.g.*, Anthony W. Flores et al., *False Positives, False Negatives and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."* 80 FED. PROB. 38, 39–40 (2016); Jason Tashea, *Courts are Using AI to Sentence Criminals. That Must Stop Now*, WIRED MAG. (Apr. 17, 2017, 7:00 AM), https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/.

[7] *See, e.g.*, Karen Hao, *AI is Sending People to Jail—and Getting it Wrong*, MIT TECH. REV. (Jan. 21, 2019), https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/.

[8] Corbett-Davies, *supra* note 4.

[9] DIETERICH, *supra* note 5, at 2–3.

[10] *See* Sorelle Friedler et al., *A Comparative Study of Fairness-Enhancing Interventions in Machine Learning*, *in* FAT*19 Proceedings of the Conference on Fairness, Accountability and Transparency 329, 329 (2019); Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, SOC. METHODS & RES. 1 (2018); Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, *in* SIGMETRICS '18 Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems 40, 40 (2018); Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 153–54 (2017); Arvind Narayanan, *FAT* 2018 Translation Tutorial: 21 Definitions of Fairness and Their Politics*, YOUTUBE (Apr. 18, 2018), https://www.youtube.com/watch?v=wqamrPkF5kk.

Building on these two key points, this paper makes two contributions. First, it suggests a typology of matching the CS notions of fairness with their appropriate legal mechanisms, it acknowledges the legal limitations of each notion, and it provides examples for potential policy domains where each notion can be implemented. Second, it outlines the practical steps that are needed to be taken by developers and policymakers in order to implement the typology and better address algorithmic fairness. The paper concludes that algorithmic fairness is contextual. Each notion of fairness can fit in a specific legal framework that makes it suitable for some policy domains more than others. In all policy domains, achieving fairness requires balancing between competing values; however, the question is where to draw the line. For example, in the context of criminal justice, algorithmic fairness means finding the right balance between public safety and individual justice; in the context of healthcare, algorithmic fairness means balancing between providing patients with the best care and general resource allocation and management considerations.[11] In those areas, balancing between the accuracy of the prediction to fairness for all is very sensitive and could have a significant impact on the life of the individual; in other areas like marketing and advertising, the balance could be subjected to a lower standard of scrutiny. Thus, the legal and social frameworks surrounding the different policy domains require adopting a different type of notion of fairness to a certain set of rules and norms. Knowing the most suitable legal mechanism for each notion of fairness would allow policymakers to choose the right tool from the toolbox, the tool that best matches the domain that algorithms are implemented in. It would also allow developers to design algorithms that are more aligned with the legal requirements.

The computer science (CS) literature refers to more than twenty different notions of fairness.[12] A large chunk of each paper addressing a new notion of fairness is devoted to arguing and explaining why the chosen notion is better than previously discussed notions, and why it is fairer, more ethical, and accurate.[13] This high number of notions represents the various attempts to address legal and social criticisms surrounding the use of artificial intelligence (AI) algorithms and different technical ways to address protected attributes, mitigate bias, and enhance explainability.[14] The CS literature has proposed different ways of grouping the 20 or so definitions of fairness and finding a common denominator between them.[15] Previous work by legal scholars who

[11] *See* Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 COLUM. L. REV. 67, 101–02 (2017) (detailing the notions behind equal group error rates as a metric of fairness regarding case recidivism rates).

[12] Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, FairWare ‘18 Proceedings of the International Workshop on Software Fairness 1, 2–3 (2018); Narayanan, *supra* note 10.

[13] Nripsuta-Ani Saxena et al., *How do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness*, AIES ‘19: Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society 1, 1–2 (2019).

[14] *See id.* at 100 (suggesting that technological awareness of public attitudes can ensure that designs are sensitive to the prevailing notions of fairness in a society).

[15] *See, e.g.*, Ninareh Mehrabi et al., *A Survey on Bias and Fairness in Machine Learning*, ARXIV 1908.09635v2, 1, 9–12 (Sept. 2019) (listing a taxonomy for fairness definitions that machine learning researchers have defined in order to avoid the existing bias AI systems); Babak Salimi et al., *Data Management*

matched computational notions of fairness to legal mechanisms has focused predominantly on the compliance of the CS notions with anti-discrimination laws, mainly through the lenses of disparate impact and disparate treatment.[16] This distinction is particularly important given the central role that disparate impact and disparate treatment play in broader doctrines such as equal protection.[17] Under the disparate treatment theory, liability could be imposed if there is an explicit classification based on the protected attribute or if there was an intent or motive to discriminate.[18] Under the theory of disparate impact, even if the policy is neutral on its face, if there is a disproportionately adverse impact on minority groups, liability will be imposed.[19] The relevant questions are: whether there is a business justification for such outcome and whether there are less discriminatory means of achieving the result.[20] This distinction has been very helpful in directing the CS community in developing mathematical solutions for algorithmic bias.[21]

This paper introduces a typology that goes beyond the disparate impact and disparate treatment analysis, in order to match the main CS notions of fairness with more concrete legal mechanisms. The notions will be divided into three categories, since each category corresponds with a different legal principle: (1) individual fairness notions—aim to achieve fairness toward the individual regardless of his/her group affiliation and corresponds with the principle of equal opportunity; (2) group fairness notions—aim to achieve fairness toward the group that the individual belongs to, and corresponds with the principle of affirmative action; and (3) causal reasoning notions—put the focus on the causal relationship between the factors and the outcome, those notions correspond with the principle of due process. The sub-notions that belong to each category will be also addressed and tied to a more concrete mechanism of the corresponding legal principle. It is important to clarify that all the notions aim to be fair toward the individual, but they differ in the weight they give to certain characteristics. Individual fairness notions focus on finding similarities between individuals and they see the individual as a whole without highlighting any particular characteristics. Group fairness notions highlight the group affiliation of the individual like gender, race, age group, etc., and the debate is about which affiliation should be considered in order to achieve the fairest outcome. Causal reasoning notions ask, for each individual, what are the factors with the highest

for Causal Algorithmic Fairness, ARXIV 1908.07924v3, 1, 1–2 (Oct. 2019); Verma & Rubin, *supra* note 12, at 3.

[16] *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 672–73 (2016) (stating that data mining will force society to rebalance the justifications for antidiscrimination law); *see also* MacCarthy, *supra* note 11, at 88 (describing the analysis of a responsible organization in turning to both recent computer science work as well as the law).

[17] Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. (forthcoming 2020).

[18] Barocas & Selbst, *supra* note 16, at 694–96.

[19] *Id.* at 701–02.

[20] *Id.* at 694–97.

[21] Muhammad Bilal Zafar et al., *Fairness Constraints: Mechanisms for Fair Classification*, 54 J. MACHINE LEARNING RES. 1, 1–2 (2017); Hoda Heidari & Andreas Kraus, *Preventing Disparate Treatment in Sequential Decision Making*, 18 PROCS. OF THE 27TH INTL. JOINT CONF. ON ARTIFICIAL INTELLIGENCE 2248, 2248–49 (2018); Michael Feldman et al., *Certifying and Removing Disparate Impact*, KDD '15 259, 259–60 (2015).

correlation with the outcome that should be included in the prediction. This division into three groups facilitates the discussion on the deeper ties between the computational notions and the legal limitations of each one.

The goal of this typology is to highlight the fact that beyond anti-discrimination, there are other legal considerations and policy measurements that could tilt the balance toward choosing one notion over the other in a certain policy domain, like looking beyond the specific case and balancing between short- and long-term policy goals. For example, even though notions of fairness that are based on affirmative action are not always the most efficient way to go computationally,[22] legally speaking, if their use is approved by the court or the legislator, they could be very useful in remedying past discrimination or to ensure a better future. This could be the case, for example, in domains like hiring and school admission. In other instances, while computationally, notions that aim to equalize the types of errors that the algorithm makes are considered to be very accurate,[23] from the legal perspective, implementing them requires determining very complicated legal and moral questions such as how much individual fairness we are willing to compromise to protect public safety. These examples are meant to illustrate that the policy domains that algorithms are operating in are too complex, and one side of the equation can be overlooked if the whole picture is not available or if we focus on solving only a very narrow segment of the social problem.

In other words, there is no one-size-fits-all solution in the context of algorithmic fairness. The typology presented in this paper acknowledges that algorithms always trade off and are often incompatible with goals of justice, and there is no single optimized fair algorithm. It aims to help policymakers to choose the most suitable notion for the case at stake, to better understand the implications of their choice; and to decide who is the appropriate person or body to make the decision and when. Therefore, in addition to the typology, the paper will highlight for developers and policymakers the practical steps needed for implementing the typology and encourage greater openness about the chosen notion of fairness. For developers, addressing fairness means communicating openly why and what specific notion of fairness has been chosen in each case, and whether other notions were examined. In addition, developers need to broaden their understanding of the complex social problem that the algorithm intends to solve, as well as the legal toolkit that is applicable for governing the specific domain. Policymakers addressing fairness are required to think deeply about the specific domains that can be automatized and altered with machine learning. On one hand, regulation by its nature is designed to be flexible and adaptable to a wide variety of cases. On the other hand, AI-based algorithms will produce the best results when the instructions given to the algorithm are as specific as possible. Thus, balancing between the two is not an easy task and it is the responsibility of policymakers to clarify the laws and policies that can be automatized, as well as providing the instructions on how to do so.

[22] Verma & Rubin, *supra* note 12, at 3.

[23] Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, 29 ADVANCES NEURAL INFO. PROCESSING SYS. 3315, 3316–17 (2016).

The paper will proceed as follows: first, the three groups (individual fairness approaches, group fairness approaches, and notions based on causal reasoning) and their sub-notions will be addressed along with their most suitable legal mechanisms. Next, in order to recap all the notions of fairness, the paper will return to the COMPAS example and use it as a case study to assess what will be the implications on pretrial procedures if we adopt each one of the fairness notions discussed in this paper. The final chapter outlines practical changes that both developers and policymakers can do in order to better address algorithmic fairness. The conclusion will follow.

The following table provides a glance into the typology that will be built in the paper. Later on, a more detailed version of this table that includes also some examples for potential policy domains that the typology can be implemented in will be used to conclude the discussion.

Table 1: Notions of fairness and summary of their corresponding legal mechanisms.

Notion	**Sub-notion**	**Corresponding Legal Mechanism**
Individual Fairness	The unaware approach	Equal opportunity as colorblindness
	Fairness through awareness	Equal opportunity based on similarities, and levels of scrutiny
Group fairness	Decoupling	Affirmative action (as separate but equal)
	Statistical or conditional parity	Affirmative action (preferably through critical diversity)
	Equal opportunity	Affirmative action (as equal opportunity)
	Equalized odds	Achieving equality by equalizing the false positive and false negative errors
	Calibration	Achieving equality by statistical significance
	Multicalibration	Achieving equality by statistical significance, and accounting for intersectionality
Causal Reasoning	Counterfactual fairness	Due process

II. INDIVIDUAL FAIRNESS AND THE CASE OF EQUAL PROTECTION

Individual fairness approaches focus on the individual regardless of his or her group affiliation. It is important to clarify for readers with technical expertise that the term “individual fairness” in this paper refers to a group of notions that evaluate the individual outside of his or her group membership. In other words, individual fairness is used here in a broader sense compared with the technical jargon that refers to individual fairness mainly as fairness through awareness.[24] It includes the notions that do not aim to achieve equality to the group of individuals that share similar characteristics; but the focus is on the individual as a whole.

From a legal perspective, individual fairness approaches go hand in hand with an individualized justice notion and equality before the law. Equality before the law is a well-established principle that has received international recognition in several documents.[25] In the United States, the principle is anchored in the Fifth and Fourteenth Amendments to the Constitution, in particular in the Equal Protection Clause, which meant to limit the ability of the government to discriminate against individuals.[26]

Individuals vary much more than groups do, thus it could be claimed that it is inaccurate to draw any conclusion about the individual based on the performance of a group of people that share similar characteristics.[27] Basing the model on group prediction can contribute to increasing the efficiency of the system, but at the expense of traditional due process safeguards.[28] Therefore, when focusing on the individual, we eliminate the fear of exacerbating traditional biases by linking between group affiliation and the individual.[29] However, as will be demonstrated below, it may be undesirable to separate an individual from his or her group affiliation and it is hard to mathematically achieve and enforce this.

A. The Unaware Approach/No Fairness Constraints

Mathematically, according to this approach, the algorithm should be blinded or unaware of any identifiable factors and prohibited attributes by law such as gender, race and sexual orientation.[30] After removing the prohibited attributes, the factors with the highest correlation to the outcome will be considered.

Legally, this model is built on the traditional way to view equal protection: color and race blind.[31] The belief is that everyone can succeed in society in proportion to their effort and talent, and any recognition of race and ethnicity is

[24] Cynthia Dwork et al., *Fairness Through Awareness*, TCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference 214, 215–17 (2012) ("In order to accomplish this individual-based fairness, we assume a distance metric that defines the similarity between the individuals. This is the source of 'awareness' in the title of this paper.").

[25] *See, e.g.*, International Covenant on Civil and Political Rights art. III, Dec. 16, 1966, S. Treaty Doc. NO. 95-20, 999 U.N.T.S. 171; International Covenant on Economic, Social and Cultural Rights art. II, Dec. 16, 1966, S. Treaty Doc. NO. 95-19, 993 U.N.T.S. 3; G.A. Res. (III) 217A, Universal Declaration of Human Rights art. VII, (Dec. 10, 1948).

[26] U.S. CONST. amend. V, amend. XIV.

[27] Sonja Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 842 (2014).

[28] Amber Marks et al., *Automatic Justice? Technology, Crime and Social Control*, *in* OXFORD HANDBOOK OF LAW, REGULATION AND TECHNOLOGY 705, 724 (R. Brownsword et al. eds., 2017).

[29] Eric Silver & Lisa L. Miller, *A Cautionary Note on the Use of Actuarial Risk Assessment Tools for Social Control, Crime and Delinquency*, 48 J. RES. CRIME & DELINQ. 138, 139–40 (2002).

[30] David Weinberger, *Playing with AI Fairness*, GOOGLE'S PEOPLE & AI RES. INITIATIVE: WHAT-IF TOOL, https://pair-code.github.io/what-if-tool/ai-fairness.html (last accessed Jan. 21, 2020).

[31] *Id*.

understood to be mere stereotyping and denying individuality.[32] The colorblind standard was quite appealing after the American Civil War since the basis for the cruel discrimination which sparked the war was explicitly color.[33] In addition, framing the discussion around colorblindness bypassed the need to determine moral questions related to race and culture, since justice and morality were assumed to be embodied in colorblindness.[34] However, to date, there is growing criticism toward the concept of individual fairness and colorblindness.[35] There is wide recognition that success is not just a matter of talent and merit, but also reflects access to resources and opportunities that are key for achieving success in our highly competitive society.[36] Legal scholars thus reached the conclusion that individual fairness is not really promoting fair outcomes;[37] the model ignores group differences, identities, and life experiences that in fact lead to different outcomes.

From a computational perspective, stripping data of protected attributes is insufficient to guarantee that race is not factored into the decision because other factors can serve as proxies to race (e.g., zip code). There are empirical studies that quantify this criticism and show the ineffectiveness of race blindness.[38] These studies show that if we insist on race blindness, we will maximize the profit of the dominant party and push people from disadvantaged groups further to the margins.[39]

The unaware approach can work only in cases where inequality is not an issue; a highly sterilized environment where the group of individuals that the algorithm classifies between is very homogenic.[40] One example of such a situation is an algorithm used for marketing a niche product for a very specific segment of the population, such as luxurious cars or assistive devices for people with disabilities. Before implementing such an approach, it is worth taking a very close look at the group of individuals we are comparing and assessing whether they are in fact equal. For example, even among the richest 10%, in marketing a luxurious car, there is probably discrimination between males and females, and certainly females of color. Thus, the population of the wealthiest 10% is not, in fact, equal and a marketing algorithm that relies on the unaware approach will still not be deemed fair.

[32] Hazel Rose Markus et al., *Colorblindness as a Barrier to Inclusion: Assimilation and Nonimmigrant Minorities*, *in* ENGAGING CULTURAL DIFFERENCES: THE MULTICULTURAL CHALLENGE IN LIBERAL DEMOCRACIES 453, 460 (Richard Shweder et al. eds., 2004).

[33] John A. Powell, *Tracing the History of Racial Inclusion and Debunking the Color-Blind/Post-Racial Myth*, AM. BAR ASS'N (Jan. 1, 2014), https://www.americanbar.org/groups/crsj/publications/human_rights_magazine_home/2014_vol_40/vol_40_no_1_50_years_later/history_racial_inclusion_color_blind_myth/.

[34] Jerome M. Culp, *Colorblind Remedies and the Intersectionality of Oppression: Policy Arguments Masquerading as Moral Claim*s, 69 N.Y.U. L. REV. 162, 163 (1994).

[35] Powell, *supra* note 33.

[36] *Id.*

[37] Neil Gotanda, *A Critique of "Our Constitution is Color-Blind,"* 44 STAN. L. REV. 1, 2–3 (1991); Randall Kennedy, *Colorblind Constitutionalism*, 82 FORDHAM L. REV. 1, 2–3 (2013); Martha Minow, *After Brown: What Would Martin Luther King Say?*, 12 LEWIS & CLARK L. REV. 599, 607–08 (2008).

[38] *See* Hardt et al., *supra* note 23, at 3321–22.

[39] *Id.*

[40] *See* Markus et al., *supra* note 32 (demonstrating how damaging a colorblind approach could be when these factors are not present).

B. Fairness Through Awareness

Still in the realm of individual fairness, this approach aims to treat similarly situated individuals similarly. The notion was coined in a prominent paper from 2011 by Professor Cynthia Dwork and her colleagues. This approach puts the focus on the individual and asks how he or she feels when someone that is as qualified as them is treated differently.[41] The similarity between individuals is defined using a mathematical metric, with its purpose being to make sure that the distance between each one of the compared individuals is similar.[42] The metric is society's best guess at the time, meaning that it can and should be updated based on developments in perceptions, beliefs, and agreements of society.[43] The metric is developed in consultation with experts in the specific domain that the algorithm is implemented in.[44] The experts educate the developers about findings that are considered state of the art, cutting edge research, and issues with a general agreement or conflicts about in the field. The developers then compute all the information and create the metric, which will be very nuanced up to the level that it will be able to determine whether a pair of individuals in the database are similarly situated. Social affiliation will be taken into account in the metric, but since it is supposed to be very nuanced, it will also be able to determine things like how much a person identifies herself with the group that we assume she is affiliated with. In other words, the metric applies more of a clinical assessment that is tailored to the individual and not an actuarial assessment.[45]

To illustrate this idea, consider the following example from the field of criminal justice: Black defendants are overrepresented in the criminal justice system.[46] They are over-policed, over-prosecuted, and over-convicted.[47] One of the factors with the highest correlation with recidivism is criminal history, and defendants with lengthier criminal histories are more likely to reoffend.[48] Therefore, any metric built for assessing the risk to recidivate will have to address this issue. Black defendants have, on average, lengthier criminal histories compared to white defendants, but since they are more frequently targeted, it would be fair to assume that some priors are for only minor crimes.[49] Imagine that, hypothetically a black defendant with five priors is equally as risky as a white defendant with three priors, such that these individuals would be similarly-situated defendants. Thus, while using the number of priors is easy to understand, this may ultimately be an unfair comparison because "priors" is a

[41] Dwork et al., *supra* note 24, at 11–12.

[42] *Id.* at 1.

[43] *Id.* at 19.

[44] *Id.*

[45] *Id*. at 1.

[46] BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING AND PUNISHING IN AN ACTUARIAL AGE 1–2 (2007).

[47] *Id.*

[48] UNITED STATES SENTENCING COMMISSION, THE PAST PREDICTS THE FUTURE: CRIMINAL HISTORY AND RECIDIVISM OF FEDERAL OFFENDERS, 6 (2017).

[49] *See* HARCOURT, *supra* note 46, at 1–2 (indicating that Black people are disproportionately likely to be charged with a crime).

very broad category, and someone might have only one prior, but it may be for something very severe, such as murder, so it would not be fair to treat him as less risky than another defendant who may have three minor priors. Therefore, the next—and more interesting—step would be to include the severity of the crime in the equation, meaning that different signals will get different weights. Similarly situated defendants will be those with not only similar numbers, but similar types of priors. Still, there are probably at least thousands of black and white defendants with a similar list of priors. Thus, after factoring in the metric of both the numbers and the severity of the crime, we will have to rely on the magic of the metric in finding small nuances that cannot be detected within the general discussion. For example, the metric can tell us that two defendants that both have two priors of the same type are not similarly situated because although both of them are black, one grew up in a very rich neighborhood, and has been given a very good reintegration plan in society. The other was raised in a very poor neighborhood, spent some time living on the street, and has been given inadequate legal representation in the past. The idea behind this example is to illustrate that the metric is allowed to take group affiliation into account, thus, proxies for race will no longer be a problem; and more attention will be given to the unique characteristics of white and black defendants.

Although group affiliation is taken into account, this approach should not be confused with group fairness notions that will be discussed in the next section. The focus of fairness through awareness is still the individual since the goal of the metric is to compare every pair of individuals in the dataset and to make sure that they are similarly situated.

The idea is that in order to maintain equality, the metric is the tool that will determine whom to compare to whom based on social determinations that will be communicated to the metric in the learning process. Given that in the creation of the metric experts from different disciplines are involved, they are responsible for making sure that the classifications that the metric suggests do not clash with equal protection as it will be now explained.

From the legal perspective, the difference between this approach and the unaware approach centers on this question: Who are we comparing the individual to when assessing if a law is discriminatory? Equal protection jurisprudence distinguishes between two schools of thought that offer different answers. According to the non-comparative justice school, we should draw a direct line between the individual and the law and assess if the treatment is discriminatory.[50] According to the comparative justice school, we cannot determine if the treatment is discriminatory without also looking at how others are treated.[51] To be more concrete, for example, in the case of determining the length of a particular sentence, retributivists belong to the non-comparative school, as they believe that the offender ought to get a punishment determined by culpability and desert theory.[52] In contrast, according to comparative justice,

[50] Deborah Hellman, *Two Concepts of Discrimination*, 102 VA. L. REV. 895, 900–01 (2016) [hereinafter Hellman II].

[51] *Id.* at 897.

[52] *Id.* at 898.

the punishment needs to be determined based on a standard of proportionality and in a way similar to how others who committed the same crime were punished.[53] While the unaware approach belongs to the non-comparative school, the notion of fairness through awareness belongs to the comparative justice approach. The fairness through awareness approach's starting point is that regulation by its nature seeks to differentiate between individuals and groups, and that equal protection does not call for formal and universal equal treatment.[54]

Another way to understand the different legal implications of the fairness through awareness approach and the unaware approach is through the distinction between formal equality, substantive equality, and the anti-stereotyping approach. The formal equality approach, also known as the anti-classification approach, requires formal equal treatment of individuals who are members of different groups; and it perceives any information linking between the individual and the protected class as perpetuating discrimination, meaning that the law should be blinded to any protected attribute.[55] On the contrary, the substantive equality approach, also known as anti-subordination approach, aims to equalize opportunities and outcomes between members of different groups, meaning that it considers the protected attribute in order to avoid perpetuating social hierarchy.[56] Although the starting point of fairness through awareness is similar to the anti-subordination approach, it goes one step forward and is more associated with the approach that is called in the literature as the anti-stereotyping approach, which requires treating individuals equally within the group and neutralizing any stereotypes associated with the protected class they belong to.[57] Hence, in the example presented above, the requirement is not to treat both black defendants equally just because they belong to the same protected group, but to eliminate any stereotypes associated with race.

Equal protection implies that similarly situated people should be treated similarly and not discriminated against.[58] The question is, what should the ground be for similarity, or in other words who are the similar people that should be treated similarly?[59] Equal protection analysis applies different levels of scrutiny based on the protected attribute.[60] The courts have developed three levels of scrutiny that apply to cases based on the sensitivity of the challenged decision. The three levels of scrutiny demonstrate different levels of suspicion

[53] *Id.* at 897–98.

[54] Michael Klarman, *An Interpretive History of Modern Equal Protection*, 90 MICH. L. REV. 213, 252 (1991).

[55] Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 540–41 (2018).

[56] *Id.* at 540–41.

[57] *Id.* at 543–44.

[58] *See Equal Protection of the Law*, BOUVIER LAW DICTIONARY (1st ed. 2012) ("Each law should treat similarly situated individuals in a similar manner.").

[59] Cass R. Sunstein, *Public Values, Private Interests and the Equal Protection Clause*, SUP. CT. REV. 127, 164 (1983).

[60] *See id.* at 137 ("The Equal Protection Clause is directed at the legality of classifications. When a classification is challenged, the first question is whether it is drawn on the basis of race or some other characteristic thought to call for 'heightened' scrutiny. . . . If there is such a classification, the statute must be invalidated unless it survives heightened scrutiny, either strict or intermediate.").

that the court will apply when examining legislation that classifies based on category. In other words, in the same manner that fairness through awareness calls for treating equally those that are similarly situated, the division to three levels of scrutiny helps in identifying the similarly situated cases that courts should treat equally.

Rationality Review/ Minimal Scrutiny

Located on one end of the spectrum, rationality review involves cases considered to be the "easiest" because they imply that the statutory classification is related to "a valid statutory purpose" and the court will view them without suspicion.[61] So long as the government action is reasonable and justifiable and not arbitrary or discriminatory, the involvement of the court will be minimal.[62] Such cases include, for example, the prerogative of the government to regulate the price of a certain product.[63] In this case, if the metric identifies criteria for classification subjected to a rationality review they could be easily included.

Strict Scrutiny

Located on the other end of the spectrum, strict scrutiny deals with the most sensitive cases, involving fundamental rights such as freedom of speech and privacy, or when a case refers to a protected attribute such as race.[64] Under strict scrutiny, the starting point is that the governmental action is not constitutional, unless the law is "narrowly tailored to achieve a compelling government interest" and there is no less restrictive alternative.[65] On its face, if the metric takes into account racial affiliation that might seem illegal. However, the use of a race as a classifying characteristic is not always constitutionally prohibited.[66] In the case of fairness through awareness, the highlight is on awareness, race is not used as a basis for discrimination, but it is one factor among many others used for the purpose of achieving a better prediction, meaning that it is a narrowly tailored use.[67]

Intermediate Scrutiny

Located in the middle of the spectrum, this category was created for cases that, on one hand, involve important matters, but on the other hand, do not require the highest protection of strict scrutiny.[68] It includes cases involving gender classification or commercial speech. When examining such cases, the court will uphold the governmental action only if it is "substantially related to an important governmental interest."[69] With analogy from the strict scrutiny

[61] Jeffrey M. Shaman, *Cracks in the Structure: The Coming Breakdown of the Levels of Scrutiny*, 45 OHIO ST. L.J. 161, 162–63 (1984).

[62] *Id.*

[63] Nebbia v. New York, 291 U.S. 502, 538–39 (1934).

[64] Shaman, *supra* note 61 at 162.

[65] Korematsu v. United States, 323 U.S. 214, 216 (1944); Shaman, *supra* note 61 at 162–63.

[66] Shaman, *supra* note 61 at 175.

[67] Jason R. Bent, *Is Affirmative Action Legal?*, 108 GEO. L.J. 55 (forthcoming 2020).

[68] Shaman, *supra* note 61 at 162–163.

[69] Craig v. Boren, 429 U.S. 190, 197 (1976); Shaman, *supra* note 61 at 163–64.

analysis, including in metric classifiers, such as gender, could be justifiable so long as race is not the only factor and it does not impose discriminatory racial classification.

Levels of Scrutiny and AI

Fairness through awareness could be legally justifiable so long as the public value that the classification is meant to serve is clear and has significant importance, given that the classification will be treated according to one of the three levels of scrutiny and in line with equal opportunity requirements. In addition, similar to the way the Supreme Court developed the three-tiered approach to judicial review, it is recommended that policymakers develop their own guidelines and hierarchy of categories to be examined under strict, middle, or minimum scrutiny. The risk that AI poses to each situation is different. The level of risk and therefore the level of scrutiny that needs to be applied should therefore be assessed carefully. There is no doubt that providing customers with financial advice or assisting judges in determining who should be jailed are very serious decisions in which the level of scrutiny that should be applied should be high.[70] On the contrary, in recommending ads or movies to watch we can tolerate lower levels of scrutiny. In the case of judicial review, the division into three levels of scrutiny is not always clear and there is an ongoing scholarly debate that calls for the creation of more scrutiny levels and application of stricter scrutiny to certain categories over others.[71] This is a healthy debate that should be encouraged in the event that levels of scrutiny for algorithmic systems will be created.

Fairness through awareness is a promising direction both mathematically and legally, and it has the potential to be implemented in highly sensitive cases such as finance and criminal justice due to its strong focus on the individual. However, researchers are still struggling to define the metric and decide on the proper categories for classifying the similarly situated individuals.[72] It is a very hard task from the policy perspective to identify the social denominators that can be computed in the metric. Controversies between scholars will always exist, but if the metric is open for review by experts, and the cohort of experts involved in building it is broad, the controversial questions could be clarified.[73]

[70] Deloitte, AI AND RISK MANAGEMENT: INNOVATING WITH CONFIDENCE 18 (2018), https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/deloitte-gx-ai-and-risk-management.pdf.

[71] Randall R. Kelso, *Standards of Review Under the Equal Protection Clause and Related Constitutional Doctrines Protecting Individual Rights: The "Base Plus Six" Model and Modern Supreme Court Practice*, 4 U. PA. J. CONST. L. 225, 226–27 (2002).

[72] *See, e.g.*, Dwork et al., *supra* note 24, at 3 ("As noted above, the metric should (ideally) capture ground truth. Justifying the availability of or access to the distance metric in various settings is one of the most challenging aspects of our framework, and in reality the metric used will most likely only be society's current best approximation to the truth.").

[73] *Id.* at 1–3.

III. GROUP FAIRNESS AND THE CASE OF AFFIRMATIVE ACTION

Mathematically, group fairness is the largest category of fairness and includes many different sub-notions. Since historical discrimination has revealed that different groups react differently to specific situations, then according to the group fairness approach, the needs and unique characteristics of each group should be addressed not individually, but collectively. Group fairness focuses on the fairness of the outcome and asks whether the outcome systematically differentiates between people who belong to different groups.[74] As will be demonstrated below, there are many different group-level outcomes one can quantify, and there is a different notion of group fairness for each of them, including mechanisms for equalizing the outcome between all the groups in question.

Legally, group fairness approaches aim to improve the position of disadvantaged minority groups in society, hence they represent different attempts to equalize certain aspects of the equation for achieving a better result.[75] Thus, group fairness notions are associated with the affirmative action mechanism.[76] Group fairness notions constitute affirmative action in the sense that they are affirmative active measures, intended to achieve actual equality or nondiscrimination, whose success metrics are defined by whether their effects are in fact more equal or nondiscriminatory.[77] They seek to achieve those outcomes but treating differently groups that in the real world are treated unequally or have unequal outcomes, by designing that differential treatment to achieve equal treatment or outcomes of the groups.

Affirmative action is a legal mechanism that aims to improve the position of historically disadvantaged minorities in the society by prioritizing them in resource allocation.[78] Affirmative action policies arose in the late 1960s after courts and governmental agencies were struggling to implement the rulings of *Brown v. Board of Education*,[79] in part because of pressure from and frustration among the black American community.[80] The set of scenarios in which courts and legislators approve of affirmative action is very narrow and mainly includes hiring, school admission, and the awarding of government contracts and licenses.[81] One example of such policy is the change in the requirements for government contractors: In addition to the standard requirement that they not

[74] David Madras, *Fairness in Machine Learning: An Overview*, U. TORONTO: MACHINE LEARNING GROUP 8 (Nov. 27, 2017), www.cs.toronto.edu/~madras/presentations/fairness-ml-uaig.pdf.

[75] *See, e.g.*, Dwork et al., *supra* note 24, at 3 ("[W]e can envision classification situations in which it is desirable to "adjust" or otherwise "make up" a metric, and use this synthesized metric as a basis for determining which pairs of individuals should be classified similarly.").

[76] *Id.* ("[The group fairness approach] is consistent with the practice, in some college admissions offices, of adding a certain number of points to SAT scores of students in disadvantaged groups.").

[77] *Id.*

[78] John Valery White, *What is Affirmative Action?*, 78 TUL. L. REV. 2117, 2119–20 (2004).

[79] Brown v. Bd. of Educ. of Topeka, Kan., 349 U.S. 294 (1955).

[80] ANDREW KULL, THE COLOR-BLIND CONSTITUTION 166–67 (1998).

[81] *Id.* at 198–99.

discriminate, they are required to develop detailed programs with concrete steps they will take to increase diversity among their workers.[82]

Successful affirmative action requires a full understanding and definition of the discrimination at play. As is the case in school admissions, giving black applicants the opportunity to sit for elite schools' admission tests did not increase the number of blacks that were accepted to the program because the tests were designed with cultural references that advantaged applicants from the dominant group.[83]

Legal scholars disagree about the effectiveness and benefits of affirmative action. While some scholars highlight the importance of affirmative action as reinforcing diversity and inclusion,[84] others believe that affirmative action is a threat to fundamental values such as fairness, equality, and democratic opportunity.[85] At the core of the disagreement is the differential treatment of individuals on the basis of merit. While for some, guaranteeing equal protection cannot mean one thing when applied to different individuals, for others, the content of the right might justify the difference and call for applying a different level of scrutiny.[86] Affirmative action has more than one dimension, and there are many legal theories of affirmative action that correspond with the different CS notions that will be discussed in this chapter.[87] The different theories center around different ways to address diversity. The colorblind diversity model and the segregated diversity model are theories that put the focus on quotas and race- or gender-based preferential treatment.[88] According to these theories, embracing cultural differences is extrinsic to the discussion about disparities in power, status, wealth and access.[89] This approach could lead to "silencing" discussions about reasons for discrimination in the first place and individuals who might criticize how affirmative action is being implemented.[90] In addition, even if the society becomes more diverse in general but at the intersection of power the majority group will still be dominant, this can strengthen the fear that only privileged individuals from the minority group will benefit from affirmative action.[91] On the contrary, critical diversity calls for linking between cultural differences and equity and parity.[92] The starting point is that resources are not distributed equally, so in order to promote diversity, we ought to dedicate special resources to traditionally disadvantaged groups in order to sustain the

[82] *Id.*

[83] Jeffrey B. Wolff, *Affirmative Action in College and Graduate School Admissions—The Effects of Hopwood and the Actions of the U.C. Board of Regents on Its Continued Existence*, 50 SMU L. REV. 627, 638–39 (1997).

[84] Elise C. Boddie, *The Future of Affirmative Action*, 130 HARV. L. REV. F. 38, 39 (2016).

[85] White, *supra* note 78, at 2119–120.

[86] Ric Simmons, *Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal Justice System*, 2016 MICH. ST. L. REV. 947, 972 (2016).

[87] *See* Cedric Herring & Loren Henderson, *From Affirmative Action to Diversity: Toward a Critical Diversity Perspective*, 38 CRITICAL SOC. 629, 630–32 (2012) (discussing the dimensions of affirmative action and a history of the meaning of diversity).

[88] *Id.* at 632.

[89] *Id.*

[90] *Id.* at 633.

[91] *Id.*

[92] *Id.* at 636.

well-being of all members of society and to help them fulfil their talents. The redistribution of resources should be done in accordance with need and equity. Critical diversity is linked with business success, because it aims to bring a new and unique talent and enhances open-mindedness, creativity, and good performance.[93]

It is important to mention that the CS group fairness notions mentioned in this section aim to achieve affirmative action but the term group is interpreted in a broader sense. Meaning that in some instances the classification is based on a certain characteristic of the individual or a protected attribute like race, gender, or age, and we would aim, for example, to achieve affirmative action in loan determination for women in general.[94] In other instances, however, the classification could focus on a certain subgroup inside the main group, for example women who are more likely to return the loan.[95] The way affirmative action will be achieved could also vary and in some instances it requires equalizing the outcome, while in other instances it requires adapting the dataset, tuning the calculations and the innerworkings of the algorithms, etc. That's what affirmative action is: treating groups that face differently discriminatory conditions outside the context differently, so as to achieve outcomes within the context that are less tainted by the discriminatory treatment those groups face outside of the context.[96]

A. *Decoupling*

The first notion of group fairness calls for classifying individuals based on their group affiliation and creating a different algorithm for each group. In other words, this approach aims to identify, for each group, the classifiers with the highest correlation with the outcome, acknowledging that the list of factors and their weight might vary across groups.[97] The advantage of this approach is that it helps in overcoming the problem of lack of data about underrepresented groups and minorities.[98] This approach distinguishes between individuals based on the sensitive attribute and avoids biases that might occur where the classifier was optimized according to the majority group but not to the minority.[99]

COMPAS, a risk assessment tool used to estimate recidivism risk, applies such an approach. COMPAS is comprised of two different algorithms, COMPAS general and COMPAS Women, which is more tailored to the needs and unique characteristics of women. COMPAS Women was created because females comprise a very small (statistically insignificant) portion of the criminal justice system,[100] so the general algorithm might miss attributes specific to

[93] *Id.* at 636–37.

[94] *Id.* at 635.

[95] *Id.* at 638.

[96] *Id.* at 631.

[97] Cynthia Dwork et al., *Decoupled Classifiers for Group-Fair and Efficient Machine Learning*, 81 PROC. MACHINE LEARNING RES. 1, 1–2 (2018) [hereinafter Dwork II].

[98] *Id.* at 1.

[99] *Id.* at 2.

[100] *See* FED. BUREAU PRISONS, INMATE STATISTICS: INMATE GENDER (Nov. 2018), https://www.bop.gov/about/statistics/statistics_inmate_gender.jsp; SENTENCING PROJECT, FACT SHEET: INCARCERATED WOMEN AND

women derived from the data. In addition, COMPAS Women is meant to be more sensitive to the specific needs of women. Thus, it takes into account economic marginalization, trauma, victimization and abuse, mental health, dysfunctional intimate relationships, self-efficacy, and parental stress.[101] The developers of COMPAS (Equivant, previously named Northpointe) do not provide sufficient information about the difference between the general version of COMPAS and COMPAS Women, and it is not clear whether factors such as marginalization and trauma are simply not considered in the general algorithm or if they are taken into account but given a different weight. Those differences can completely change the outcome and impact fairness. COMPAS developers claim that integrating gender sensitivity into the risk assessment tool will help agencies to achieve fairer results. It is important to mention in this context that a State Supreme Court case that analyzed the reliance of COMPAS on gender ruled that the use of gender by COMPAS is non-discriminatory but the opposite, its purpose is to increase the accuracy of the prediction; and in addition, this does not violate defendants' right to due process.[102]

The Challenges of This Approach

First, while creating different algorithms for different groups might put disadvantaged groups in a better position, a remaining open question is whether we as a society are willing to allow this. In the same manner that we have COMPAS Men and COMPAS Women, is it acceptable to have also COMPAS White and COMPAS Black? Can we create a different risk assessment tool for transgender or homosexual individuals to integrate their unique needs? These are tricky questions, and one fear is that having separate algorithms for separate groups might give the impression that we are back to the separate but equal era. It is no secret that separate but equal is not equal. True equality was not the main motivation behind the Jim Crow laws enacted in the late 1800s and early 1900s. The segregation, especially when enforced by the law, showed superiority of whites and inferiority of blacks which had a detrimental effect on the quality and level of services and goods provided to blacks in the separate premises.[103] Early on, the doctrine was used to reinforce racial segregation and discrimination. Decoupling takes a different approach; the idea is to use a similar doctrine in order to reduce disparities between disadvantaged and advantaged groups. Decoupling implies that in some cases where society is already divided, there is no formula that works identically for all groups, so separation is necessary in order to achieve equality. Nevertheless, when the classification is based on a sensitive attribute such as race, the stamp itself is iniquitous.[104] Flagging the black community as a community that needs to be

GIRLS (2018), https://www.sentencingproject.org/wp-content/uploads/2016/02/Incarcerated-Women-and-Girls.pdf (noting in the federal system, 7% of all inmates are female, and in the state system, although the number of women is constantly growing, they still make up only a very small fraction of the total correctional population).

[101] *COMPAS WOMEN*, NORTHPOINTE, http://www.northpointeinc.com/files/downloads/Womens.pdf.

[102] State v. Loomis, 881 N.W.2d 749, 767 (Wis. 2016).

[103] KULL, *supra* note 80, at 165.

[104] *Id.* at 103.

separated from the rest of the population using the excuse of equality could be very dangerous. Even after the rulings in *Brown*, it took many years to end the segregation in practice, and federal troops had to intervene in several cases to ensure that the ruling was implemented. Classification based on gender is also sensitive, but in the case of COMPAS Women, the fact that women make up a truly small portion of the criminal justice system is the justification for such practice. The fact that black defendants are overrepresented in the criminal justice is already causing prejudice, so creating a separate algorithm for scoring their risk is likely to increase the disparity.

Second, intersectionality and enumeration could pose a significant challenge to this approach. Individuals could belong to more than one group, or one might be discriminated against on several grounds. Therefore, it is not always clear how to define the subgroups, or how far to subdivide, since the information might be obscured.[105]

Third, some legal scholars assert that using race explicitly in the algorithm would constitute a disparate treatment on the basis of race and therefore it is prohibited to include in the algorithm a feature that its predictive power is different across groups.[106] However, the answer to the question whether using race as a determinative factor for a separate algorithm would vary based on the context.[107] But if the algorithm is using different but relevant information in order to evaluate each group similarly, this does not constitute disparate treatment.[108] For example, if housing stability is a good predictor for recidivism for white defendants but not for black defendants, not including this criteria in the algorithm that evaluates black defendants will improve fairness and accuracy without causing a disparate treatment.[109] In any case, a social debate about the impact of such separation needs to take place.

Given the limitations, decoupling as a notion of fairness will be most useful only in a very particular set of cases where we can identify a specific minority group that is very small and unique.[110] The idea is that, without a separate classifier, the needs and distinctive characteristics of this minority group will be ignored.

It is very important to mention that in this section, decoupling has been discussed as a separate notion of fairness that creates different classifiers for different groups. However, decoupling is in fact a broader concept, and some sort of explicit or implicit decoupling could be embodied in many other notions of group fairness.[111] So the limitations of this approach should also be taken into account when considering other notions. In practice, the classifiers themselves can be designed to enforce other group notions of fairness since other notions often create different thresholds for different groups in order to equalize

[105] Dwork II, *supra* note 97, at 3.
[106] Hellman, *supra* note 17, at 38–39.
[107] *Id.*
[108] *Id.* at 40.
[109] *Id.* at 38–39.
[110] Dwork II, *supra* note 97, at 2.
[111] *Id.*

the true positives, true negatives, false positives, or false negatives.[112] Also, a complex algorithm could be performing a form of implicit decoupling if it is given some information that correlates with sensitive attributes.[113] For example, an algorithm given the zip code of defendants could combine the inputs differently based on zip code, such that white and black defendants are, on average, subject to a slightly different classifier.[114] In other words, there is a difference between the definition of fairness and its actual implementation, and this could have a direct impact from a legal or policy perspective. On the one hand, if it is not legal to subject different groups to different classifiers and de facto this is what is happening, this needs to be addressed and may be hard to overcome. On the other hand, there might be a benefit in declaring that everybody is subject to the same algorithm and some adjustments are made under the hood to accommodate group differences if those notions are effective.[115]

B. Statistical Parity and Conditional Statistical Parity

Statistical parity, sometimes called demographic parity, aims to ensure that the fraction of people from group A who receive a particular outcome is the same as the fraction of group A of the whole population.[116] For example, since women make up 50% of the U.S. population, 50% of granted loans should be given to women.[117] Or in the context of the criminal justice system, if blacks make up approximately 15% of the American population, 15% of all those who will be classified as low-risk or high-risk should be black defendants.[118]

Conditional statistical parity is a unique form of statistical parity.[119] It aims to equalize the outcomes between different groups, conditioned on some factors.[120] For example, if the factors that a credit-scoring algorithm considers are the requested loan amount, employment, and age, conditional statistical parity would make sure that both men and women between twenty and thirty years of age would get loans at the same rate, or that an employed female and employed male will be assigned the same probabilities of repaying a loan.[121] In criminal justice, if an algorithm that predicts the risk to recidivate considers, for example, the number of priors, socio-economic status, and age, conditional statistical parity will make sure that among black and white defendants with two priors the score will be distributed equally.[122]

[112] Hellman II, *supra* note 50, at 38.

[113] Dwork II, *supra* note 97, at 1–2.

[114] *Id.* at 2.

[115] Alex Beutel et al., *Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements*, *in* AAI/ACM Conference on Artificial Intelligence, ETHICS AND SOCIETY 453, 455 (2019).

[116] *Id.*

[117] Verma & Rubin, *supra* note 12, at 3.

[118] Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, *in* Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 797, 798 (2017).

[119] Verma & Rubin, *supra* note 12, at 3.

[120] *Id.*

[121] *Id.*

[122] *See* Corbett-Davies et al., *supra* note 118, at 798 (describing how to define algorithmic fairness).

The Challenges of These Approaches

Statistical parity and conditional statistical parity are often dismissed by computer scientists as valid notions of fairness due to the following reasons.

First, legally statistical parity and conditional statistical parity calls for affirmative action that is more focused on the outcome and less on the process that led to the outcome. The common perception is that affirmative action based equal outcome calls for numerical quotas and reverse discrimination of minorities at the expense of qualified members from the majority group.[123] However, this view is both inaccurate and unconstitutional.[124] Rather, affirmative action institutions, also those that focus on the outcome, aim to encourage diversity.[125] They have the intention to break down racial and gender class hierarchies by treating people formally unequally to achieve real-world equality.[126] Maintaining successful affirmative action at the stage of the outcome is a very complex task that requires both the commitment of all stakeholders involved and the flexibility to internalize different conceptual goals of different disciplines.[127] If required to implement affirmative action in hiring, a certain company could, for example, hire the best candidates in the majority group and the worst candidates in the minority group, such that they will be able to fire them in a short amount of time. The company could also pose unrealistic working conditions that will make employees from the minority group quit, or employers could decide to hire only those from the minority group who are unlikely to accept the job. In other words, there are many things that an employer could do to sabotage affirmative action efforts, and those deeds could have a long-term effect that will broaden the gap between the majority group and minorities, and reinforce traditional biases.[128] It is very important to select the most qualified applicants in the pool in order to increase the chance of success of affirmative action, however, in cases where there are no sufficiently qualified applicants in the pool, affirmative action urges employers to hire anyway because it may help in achieving long-term goals.[129] Balancing between the short-term goals and the long-term goals, and computing this balance, is not easy given that each stakeholder has a completely different agenda.[130] To sum, it is not clear what has to be done in order to make affirmative action work in practice, and the variety of options might limit its success.[131]

[123] Sherri L. Wallace & Marcus D. Allen, *Affirmative Action Debates in American Government Introductory Textbooks*, 47 J. BLACK STUD. 659, 661 (2016).

[124] *See id.* (relaying that these myths on affirmative actions have consequences).

[125] *See* Herring & Henderson, *supra* note 87, at 639 (discussing how affirmative action attempts to assist in diversity).

[126] *Id.*

[127] *See generally* Uduak Archibong & Phyllis W. Sharps, *A Comparative Analysis of Affirmative Action in the United Kingdom and United States*, 3 J. PSYCHOL. ISSUES ORGANIZATIONAL CULTURE 28 (2013) (discussing the complexity and difficulty of addressing discrimination).

[128] *See id.* at 43 (discussing the revolving door problem without resources to continue the programs).

[129] *See id.* at 33 (emphasizing that promoting qualified candidates is essential).

[130] *See id.* at 45 (describing how the policy implemented to improve equal opportunity must be meaningful to the employer).

[131] Dwork et al., *supra* note 24, at 218–20.

Second, both statistical parity and conditional statistical parity call for equalizing the outcome without considering the reasons that led to this outcome. This approach ignores the fact that historical data might be biased, that the base rate of an outcome may be higher in certain groups, and that not all groups are equally represented in the system.[132] Thus, in order to satisfy this notion of fairness, the algorithm will have to set different thresholds for each protected group; and in a certain domain this could legally be considered disparate treatment, for example, to decide that white defendants with a score of six will be considered high risk, while black defendants only above the score of eight will be considered high risk.[133] In other domains, like school admission, we might temporarily accept the different thresholds because this is the only way to plant change in a broken and discriminatory system.

Third, it is not clear in regard to statistical parity what the overall population that the fraction of the minority group should be equalized to is; whether it is their fraction in the national population, state–based population, the applicant pool, or so on. Choosing a different definition will change the outcome significantly.[134] For example, choosing one practice over the other might cause a situation like congressional redistricting, where institutions will divide the population in a way that will maximize their interest as opposed to actually promoting a fair outcome.[135] Or if the total population is used as the number of applicants, we have to take into account that minority groups who have already suffered from the treatment of the system, lack trust in it, and therefore they apply less so the number of male and female applicants will not be equal.[136] Thus, we will end up giving loans to many men and fewer women, believing that we are satisfying statistical parity when in fact we are increasing disparities.[137]

Fourth, affirmative action is a civil mechanism, and the set of cases in which statistical parity will be implementable is quite narrow. This section uses the criminal justice system as an example of a case where applying statistical parity is complex.[138] Law enforcement is an individual act; thus, formal equality before the law means that everyone who commits a crime should be punished, and for the same crime, people should be punished in a similar way.[139] But since racial discrimination permeates every aspect of the criminal justice system, some researchers are trying to advocate for affirmative action in order to reduce the

[132] *Id.* at 225.

[133] Hellman, *supra* note 17, at 38.

[134] Dwork et al., *supra* note 24, at 215.

[135] Pamela S. Karlan, *Easing the Spring: Strict Scrutiny and Affirmative Action After the Redistricting Cases*, 43 WM. & MARY L. REV. 1569, 1575–76 (2002).

[136] *See* Verma & Rubin, *supra* note 12, at 7 (indicating a disparity between the treatment of male and female applicants).

[137] *See id.* at 2 (describing the outcomes for similar males and females).

[138] *See* Hellman, *supra* note 17, at 31 (describing some of the issues in the algorithm in the criminal justice context).

[139] *See* Paul Butler, *Affirmative Action and the Criminal Law*, 68 U. COLO. L. REV. 841, 857 (1997) (describing the tension between formal equality and unfair sentencing).

involvement of blacks in criminal justice.[140] At the core of the suggested solution is to strike a different balance between the various goals of the criminal justice system and to move from a retributive justice based system to a more restorative justice-based system.[141] Such change was implemented in Canada with the enactment of Bill C-41 in 1995.[142] One restorative justice element in the bill includes giving judges the option of conditional sentences, meaning, the option to allow defendants who were sentenced to less than two years in prison to serve their sentence in the community in order to be held accountable to the victim and the community and to focus on repairing the injury.[143] In addition, in order to reduce the high number of aboriginal defendants in prison, the bill directed judges to take into account the particular factors that might have played a role in bringing the specific offender in front of the court, and to consider all types of sentencing procedures and sanctions, including alternative restorative justice options, which might be more appropriate for the aboriginal defendant.[144]

Other affirmative action inspired solutions are to use rehabilitation as the primary criminal justice goal when deciding the length of the sentence for a black defendant as opposed to retribution, or to arrest black defendants for drug possession offenses only in proportion to their percentage in society.[145]

Despite the challenges, from a legal and policy perspective, statistical parity is a relatively easy notion to implement because influencing the outcome will have a relatively immediate impact; thus, it can be utilized for implementing policies.[146] In fact, statistical parity is being used in practice in domains such as hiring, school admission, and government contracting.[147]

The following example is meant to illustrate that even if statistical parity and conditional statistical parity are not perfect notions, they should be taken seriously by computer scientists and policymakers in areas where it is legally possible because this notion allows for a good balance between short and long term policy goals and achieving a real change in individuals' lives.[148] Take the case of a manager of an information technology (IT) company hiring software engineers. Since the conditions of the job are quite appealing and given that positions for software engineering are quite competitive, the manager decides to

[140] *See* May Lydia Yeh, *Restorative Justice, Affirmative Action Sentencing Legislation and the Canucks: Lessons from our Northern Neighbor*, 7 WASH. U. GLOBAL STUD. L. REV. 661, 664 (2008) (explaining the "war on drugs" racial disparity); Mari J. Matsuda, *Crime and Affirmative Action*, 1 J. GENDER RACE & JUST. 309, 311–12 (1998) (demonstrating that affirmative action is necessary to address inequality in the criminal justice system); *id.* at 842 (illustrating the barriers minorities face in society); Kennedy Green, *Restorative Justice*, CLAREMONT J.L. & PUB. POL'Y (Aug. 30, 2016) https://5clpp.com/2016/08/30/restorative-justice/ (considering racism within a restorative justice program).

[141] *See* Yeh, *supra* note 140, at 673 (describing growing influence of restorative justice).

[142] *Id.* at 674.

[143] *Id.*

[144] *Id.* at 674–75.

[145] Butler, *supra* note 139, at 877.

[146] *See* Dwork et al., *supra* note 24, at 217 (indicating adjusting outcomes are simple with the statistical models used).

[147] *See* Fisher v. Univ. of Tex. at Austin, 570 U.S. 297, 304–06 (2016) (describing factors the University of Texas considers in admissions); United Steelworkers of Am. v. Weber, 443 U.S. 193, 197–99 (1979) (explaining the negotiated affirmative action "bargain").

[148] *See* MacCarthy, *supra* note 11, at 110 (illustrating an example of statistical parity).

use an algorithm in helping him narrow down the number of applicants who will be invited for an interview. The manager is aware of the fact that most software engineers that are working in the company are men and fearing reinforcing a bias against women, the manager is trying to find an algorithm that is sensitive to fairness. Critics might claim that an algorithm that is based on statistical parity will not produce the best results because the algorithm will try to select the most qualified women, but some of them might still be less qualified than men in order to satisfy affirmative action.[149] However, it can be argued that statistical parity in general, and conditional statistical parity in particular, can increase fairness toward women both in the short and long term.[150] Research has shown that there are many reasons why women are underrepresented in highly competitive IT-related jobs.[151] Women are six times less likely to receive ads about high-paying jobs; thus, they have to put more effort into finding such positions.[152] In addition, the few ads that women manage to see are often worded in very masculine terms, sometimes in a way hostile toward women, which makes women less interested in applying even if they are qualified for the job.[153] Therefore, in many instances, the real problem is not that there are not enough qualified women to fill the vacancies, but that employers are not broadening the search process or adjusting the traditional language and attitudes to accommodate more women. Given that it is in the employer's interest to hire the best candidates, putting the burden on the employer to satisfy statistical parity might push them to find creative solutions and deal with the factors that caused the discrimination beforehand.

The most beneficial way to apply statistical parity will be through the lenses of affirmative action as critical diversity. It would be worthwhile to contrast the outcomes of a statistical parity based algorithm and a conditional statistical parity algorithm in the context of hiring in order to determine which one is preferable. It is likely that in some cases conditional statistical parity will be preferable, since it allows for a higher degree of compatibility to the job.[154] For instance, by considering level of education as a factor, conditional statistical parity ensures that the women who will be accepted will likely be those with relatively more education. However, in other circumstances, conditional statistical parity might pose too many limiting factors. For example, if we set the bar at ten years of experience or more, and further condition compatibility

[149] *See id.* at 112–13 (explaining the judgement of statistical parity is that less qualified individuals will be selected).

[150] *See* Verma & Rubin, *supra* note 12, at 3 (describing the results in the conditional statistical parity study).

[151] Kasee Bailey, *The State of Women in Tech 2019*, DREAMHOST (Mar. 7, 2019), https://www.dreamhost.com/blog/state-of-women-in-tech/.

[152] Samuel Gibbs, *Women Less Likely to be Shown Ads for High Paid Jobs on Google, Study Shows*, GUARDIAN (July 8, 2015), https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study.

[153] *See* Danielle Gaucher et al., *Evidence that Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality*, 101 J. PERSONALITY & SOC. PSYCHOL. 109, 109–110 (2011) (describing the bias in advertisements as well as the obstacles women face in the job market).

[154] *See* Verma & Rubin, *supra* note 12, at 3 (describing the better results for women under conditional statistical parity).

based on personal characteristics such as charisma, leadership, and management, we run the risk of excluding females because those characteristics are more associated with men. In such cases, general statistical parity is favorable for achieving our overall goal of reducing disparities. In order to decide whether to adopt a statistical parity or conditional statistical parity based algorithm, particular attention needs to be given to the factors that will be considered by the algorithm and whether or not they have the potential to discriminate.[155] It is important to mention that this approach does not rely on the altruistic motives of the potential employer, but on the realization that change is a lengthy and complicated process.[156] Thus, if employers will initially hire women and fire them after a couple of months, "forcing" them to do so year after year will increase the chances that they will invest in long-term solutions and accept women as part of the workforce.

A similar analogy can be made to school admissions. Indeed, with schools it might be an even "easier" case because schools, from the start, should be more invested in the pillars of affirmative action: enhancing diversity, promoting education for all, and providing students with the best tools for success. It is possible that requiring schools and universities to admit a certain number of students from diverse backgrounds will urge them to take the necessary steps for identifying the right candidates, such as adjusting their admission tests, enlarging their recruitment outreach, and providing ongoing support to admitted students who might be at risk to withdraw from the program.[157]

Lastly, a recent study that examined the public perception toward algorithmic fairness found that the majority of people favor statistical parity over more complicated notions of fairness even when the stakes are high.[158] Public perception is perhaps not the only factor that needs to be taken into account, but because algorithms inherently impact people's lives, developing policies that respect people's perceptions of fairness will increase the chances that algorithms will be adopted and trusted.

C. Equal Opportunity

According to this approach, the group that should be favored is the group of individuals who belong to the positive class. "Individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome,"[159] meaning that, in the loan example, both men and women who in reality will return the loan should have equal opportunity to actually receive

[155] *See* MacCarthy, *supra* note 11, at 94 (describing how in "algorithmic fairness," it is important to look at non-discriminatory classifiers).

[156] *See* Archibond & Sharps, *supra* note 127, at 32 (describing how Title VII implementation allows for affirmative action in employment).

[157] *Cf. id.* (showing how Title VII leads to better outcomes).

[158] Megha Srivastava et al., *Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning*, in Proceedings of the 25th SIGKDD International Conference on Knowledge Discovery & Data Mining 2459, 2459–60 (2019); *See* Saxena et al., *supra* note 13, at 2 (describing fairness).

[159] Moritz Hardt, *Equality of Opportunity in Machine Learning*, GOOGLE AI BLOG (Oct. 7, 2016), https://ai.googleblog.com/2016/10/equality-of-opportunity-in-machine.html.

it.[160] Similarly, both low-risk black and white defendants should have equal opportunity to be released at the same rate.[161] In other words, equal opportunity calls for giving the group that truly belong to the positive class the chance to be classified as so and makes sure that this group represents all members of society.

This approach goes hand-in-hand with a growing number of researchers who call for using risk assessment tools for identifying only low-risk defendants.[162] In such a case the tool will help in increasing the positive outcome (release), and the task of sending high-risk defendants to jail will be kept in the hands of the judge.[163] The challenge of this approach is that it shifts the focus from the hardest and most controversial cases to the "easy cases."[164] The term "easy cases" might not be the best fit for criminal justice since jeopardizing liberty is always at stake. However, a criticism of this approach is that it ignores the most vulnerable segment of society, those who are classified as high-risk, and any potential disparities within such a classification.[165]

From the legal perspective, as its name implies, this approach is associated with affirmative action as equal opportunity, in this case, an opportunity to be classified to the positive class for those who truly belong to this group.[166] While theories that are based on equality of outcome focus on the end point and demand equality directly in the outset, equal opportunity theories focus on providing an equal ground and allocating resources for filling any gap in the starting point.[167] The fairness notion equality of opportunity aims to favor in resource allocation those who belong to the positive group, meaning those who are most qualified to be included in the positive group.[168]

D. Equalized Odds

Equalized odds, sometimes called equal accuracy, has two conditions: (1) to satisfy equal opportunity; and (2) to achieve equal opportunity in the negative class.[169] Equalized odds is essentially saying that we want equal opportunity for all truly positive people to be classified as positive and equal negative opportunity for all truly negative people to be classified as negative.[170] In other words, this approach tries to equalize false positives and false negatives. It ensures that the errors that the algorithm might make are equal across all groups, and that the percentage of incorrect classifications is also equal.[171] By

[160] *See* Weinberger, *supra* note 30 (describing a demographic parity situation).

[161] *Cf. id* (explaining that in the equal opportunity situation the same amount of men and women should receive loans).

[162] Matt DeLisi et al., *Inverting Risk Assessment: Considering the Lowest Risk Clients in the Federal Criminal Justice System*, 30 CRIM. JUST. POL'Y REV. 1043, 1046–47 (2018).

[163] *Id.*; Seena Fazel et al., *Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24,827 People: Systematic Review and Meta-Analysis,* BMJ (2012).

[164] DeLisi et al., *supra* note 162.

[165] Weinberger, *supra* note 30.

[166] Wallace & Allen, *supra* note 123, at 662.

[167] *Id.*

[168] *Id.*

[169] Verma & Rubin, *supra* note 12, at 4.

[170] *Id.*

[171] Hardt et al., *supra* note 23, at 3316–17.

analogy to the examples discussed above, if the credit scoring and loan determination algorithm has a 0.85 statistical significance, it is considered to be an algorithm with a high accuracy rate.[172] Still, the algorithm is expected to wrongly classify 15% of the cases. Among those 15% there will be people who got a loan despite the fact that they will not be able to repay it, as well as people who did not qualify for a loan although they are able to repay it on time. Equalized odds explicitly holds that both the false positive rate and the false negative rate should be the same for men as for women. In the criminal justice context, equalized odds will ensure that not all of those who are wrongly sent to jail are black and not all the defendants who are released and recidivated are white. This is an important approach, but it is not easy to achieve or to agree on the way to enforce it due to the following caveats that need to be taken into account.

First, the challenge with equalizing false positives and false negatives is that society values them differently, and their economic cost is also different.[173] Setting the threshold and deciding on an error rate our society is willing to tolerate is not an easy task.[174] In some cases, false negatives will be more expensive than false positives. Take, for example, the case of testing for breast cancer. While a false positive may cause extra stress, a false negative could potentially lead to the loss of life. Still, this example should be examined cautiously because even in such a case, there is still a balance between the two. It certainly is not useful for everyone to test for breast cancer all the time to make sure absolutely no false negatives occur. In other cases, false positives could be more costly; an individual will suffer more if the algorithm incorrectly classifies him or her as high risk.[175] Spending unnecessary time in jail, for instance, is very costly for the individual and will have a direct impact on his or her current case, any future cases, and any attempt to reintegrate into society.[176] In addition, a presumption against false positives and focus on the individual is inherent in legal standards such as the Due Process Clause, the presumption of innocence, and the "beyond a reasonable doubt" standard at the center of our criminal justice system.[177] This is demonstrated by the well-known maxim: "Better that ten guilty men go free than that one innocent man be punished." These well-established legal principles, therefore, set a high burden on the government to prove the defendant's guilt. But again, individual justice should always be balanced with public safety, as the key for increasing overall fairness is achieving the right balance between public safety and individual justice.[178] Lastly, unequal odds and no other notion of fairness considers who is in the best position to bear the burden of the law. In some instances, we would explicitly put the burden of the law on the stronger party in order to achieve certain societal

[172] *Id.*

[173] Hellman, *supra* note 17, at 17.

[174] *Id.*

[175] *See id.* (measuring the fairness of algorithms).

[176] *Id.*

[177] *Id.*

[178] Adam Crawford, *Governing Through Anti-Social Behaviour: Regulatory Challenges to Criminal Justice*, 49 BRITISH J. OF CRIMINOLOGY 810, 811 (2009).

goals.[179] Hence, when a person is denied a loan for which they could have repaid, that person bears the burden—they can't buy a house or a car which could impact other opportunities related to jobs, education, etc. When a person cannot repay a loan, the lender bears the risk of nonrepayment, and it means they will make a marginally smaller return on investment, given a portfolio of loans. Balancing between false positives and false negatives will vary in severity depending on context and translating this balance to a numeric error rate is a combined technical and policy task that needs to be undertaken after consulting with all relevant stakeholders.[180]

Second, when equalizing false positives and false negatives, we do not have access to the counterfactual outcome that would have resulted if we had assigned someone the positive outcome instead of the negative, and vice versa.[181] In such cases, it becomes difficult to estimate false negative and positive rates.[182] It is impossible to know whether imprisoned individuals would have committed a crime if they had been released instead, if those who did not get a loan would have defaulted, or what would have happened to a patient who did not receive a certain treatment, if they had received it.[183] The difficulty of estimating counterfactual outcomes, details about how data was collected, and how false positive and negative rates were measured must be taken into account when practitioners attempt to enforce fairness through balancing of false positives and false negatives.

Third, it is not possible to simultaneously equalize the error rates and ensure that the algorithm is calibrated.[184] In other words, when equalizing the error rates, we will not be able to be sure that the positive predictive value is also equal; and this might lead policymakers to not trust the algorithm. The meaning and importance of calibration will be discussed next, and this limitation will be addressed again for better clarification.

E. Calibration

An algorithm that has been calibrated is an algorithm that achieved equality within any given score category that it creates.[185] An inherent aspect of algorithmic tools is calibration. Calibration in practice means that probabilities should carry semantic meaning. If we look at the set of people who receive a predicted probability of *p* for being positive in reality, we would like a *p* fraction of the members of this set to actually be positive instances of the classification problem, and vice versa.[186] For example, if there are one hundred defendants in our dataset and the classifier assigns everyone in the dataset a 0.6 probability of

[179] *Id.*

[180] Hellman, *supra* note 17, at 17.

[181] *Id.*

[182] *Id.*

[183] Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237, 261 (2017).

[184] Kleinberg et al., *supra* note 10, at 42.

[185] Geoff Pleiss et al., *On Fairness and Calibration*, 30 ADVANCES NEUTRAL INFO. PROCESSING SYS. 5680, 5680–81 (2017).

[186] *Id.*

belonging to the positive class, we expect sixty defendants to ultimately be in the positive class in reality, and that among those sixty defendants the percentage of blacks and whites is equalized.[187] Calibration is important for ensuring that decision makers have confidence in the prediction and that they will treat people who belong to the same class similarly.[188] Imagine that for a certain risk level, the algorithm predicted a general probability of 0.8 to recidivate, but when looking deeper into the data, we discover that among those who were flagged with this probability, 95% are black and only 5% are white. It is important to clarify that this is not necessarily an error rate; all those who were flagged at this risk level could in fact be risky. But the algorithm is not calibrated for different groups, so this is the reason for the difference. A judge that receives such result would probably release any white defendant that is ranked in this category because they are aware of the fact that the vast majority of defendants ranked in this category are black, so the probability that a white defendant in this category would recidivate is low. Lack of calibration would mean that black and white defendants will have different aggregated probabilities of belonging to this class. In other words, calibration justifies treating people with the same score comparably with respect to the actual outcome.[189] In addition, research that examined the public perception toward algorithmic notions of fairness found that calibration as a method is perceived as especially significant in the public eye.[190]

Calibration seems appealing as a notion of fairness; however, one complication in its use is in the way calibration interacts with other notions of fairness, especially equalized odds.[191] Mathematically it is impossible to achieve calibration, equal false positive rates, and equal false negative rates between two groups, unless you have a perfect classifier, or the two groups in question have the same base rate for the outcome.[192] For example, imagine that in a certain clinic the satisfaction rate for the doctors is approximately 80%. The algorithm that yielded this estimation rate is not calibrated, but the false positive and false negative rates are equalized. Assume that you are a patient who is trying to pick a doctor, so you dig a bit deeper in the data and discover that the satisfaction rate for the female doctors is actually 95% and 60% for male doctors. Naturally you will pick a female doctor given the huge difference between the groups. Thus, an algorithm that is not calibrated could lead policymakers to discriminate based on protected attributes, or at least to take them into account even if they are not supposed to.

In some policy domains we will prefer calibration, which means we will have to give up the goal of equalizing false positives and false negatives.[193] In those cases, we might want to use an auditing method to check that the results are not biased and that a certain type of error does not occur more than others.

[187] *Id.*

[188] Hellman, *supra* note 17, at 22.

[189] Kleinberg et al., *supra* note 10, at 43.

[190] Saxena et al., *supra* note 13, at 6.

[191] Pleiss et al., *supra* note 185, at 5681.

[192] *Id.*

[193] *Id.*

In other policy domains, the errors that the algorithm might make would be especially important to avoid, so we would choose equalized odds.

Multicalibration is a new approach in computer science that attempts to achieve a better balance between individual fairness and group fairness; and it takes intersectionality into account.[194] It defines calibration as the metric for fairness, and it does not attempt to deal with false positive and negative rates.[195] It does allow for the creation of slightly different classifiers and different thresholds for different sub-groups (some sort of decoupling), but it does not create these thresholds for false positives and negatives.[196] Instead, the main insight of multicalibration is that we might not be able to tell you which groups and intersectional groups are important to consider from day one.[197] Multicalibration does not require you to say, for example, that Hispanic women often get inaccurate predictions, and to therefore make sure that we are calibrated on them. Instead, multicalibration automatically notices that you are not calibrated on that intersectional group and tries to improve its calibration, but there is no guarantee about the equalization of false positive or negative rates.[198] The challenges of multicalibration are to therefore identify and compute all potential intersectionality and to carefully generalize from the small learning dataset to the real data. In addition, multicalibration does not deal with the problem of lack of data about underrepresented groups in the dataset, so if one of the sub-groups is too small it is not guaranteed that it will be calibrated. Lastly, the choices of the features considered by the algorithm affects which sub-groups can be identified.[199] For example, if marital status is not included as a feature, it might be hard for the algorithm to ensure fairness for married versus divorced versus single people. Multicalibration is a new approach that will be interesting to observe in its development and application to new case studies.

IV. CAUSAL REASONING AND DUE PROCESS

Causal reasoning approaches focus on the causal relationship between the factors and the outcome, meaning that not all factors with a high correlation will be immediately included, only those that have been proven or agreed upon by experts to actually cause the outcome.[200] Supporters of causal reasoning approaches believe that the traditional group fairness approaches based on observational criteria are unable to determine if disparate impact will be eliminated even after "removing" all predicted attributes and proxies.[201]

[194] Ursula Hébert-Jonson et al., *Multicalibration: Calibration for the (Computationally-Identifiable) Masses*, 80 PROCS. OF THE 35TH INTL. CONF. ON MACHINE LEARNING 1939, 1940 (2018).

[195] *Id.*

[196] *Id.*

[197] *Id.*

[198] *Id.* at 1945.

[199] *Id.*

[200] JOSHUA LOFTUS ET AL., CAUSUAL REASONING FOR ALGORITHMIC FAIRNESS, *available at https://arxiv.org/abs/1805.05859* (May 15, 2018).

[201] Niki Kilbertus et al., *Avoiding Discrimination Through Causal Reasoning*, *in* Proceedings of the 31st International Conference on Neural Information Processing Systems 656, 657–58 (2017).

Mathematically, causal reasoning-based approaches aim to achieve counterfactual fairness.[202] Counterfactual fairness identifies the factors that can cause discrimination and attempt to single out the effect of such factors.[203] They do so by creating a counterfactual world in which the individual belongs to the dominant group.[204] Since it is a very complex task to mitigate all the proxies that can cause bias, this approach addresses all the relationships between the attributes and intervenes by assigning different values to each attribute.[205] For example, it asks what would happen if a black defendant was a white defendant in the counterfactual world, or if a female job applicant was a man with the same characteristics and qualifications.

Legally, researchers claim that counterfactual causal reasoning approaches are the most common in law and social science disciplines for detecting racial discrimination.[206] This is because, for example, researchers are often asked to provide empirical evidence demonstrating what would happen to a nonwhite individual in a circumstance if they were white in order to measure the causal effect of race on discrimination, in other words if there is disparate impact.[207] In addition, courts interpret the constitutional equal protection requirement as treating similarly situated individuals similarly.[208] Thus, in order to show that a particular practice is discriminatory, the litigant has to prove an intent to discriminate or discriminatory impact.[209] In other words, that a similarly situated individual of another race or ethnicity should have been, but was not subject to the same treatment as the defendant.[210]

A. The Advantages of Causal Reasoning-Based Approaches

Relying mainly on correlations could be misleading: one of the strengths of machine learning techniques is their ability to learn from examples, generalize, and find patterns in the data.[211] The patterns are correlations between the factors fed into the machine and the outcome.[212] But correlation does not imply causation, meaning that there is no causal relationship between the factors that the algorithm identified and the outcome.[213] Supporters of implementing machine learning algorithms in our daily life claim that an algorithm based on

[202] LOFTUS ET AL., *supra* note 200.

[203] Matt J. Kusner et al., 'Counterfactual Fairness' *in* Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS, Long Beach, CA, USA, 4069 (2017).

[204] *Id.*

[205] *Id.* at 4071–74.

[206] Issa Kohler-Housman, *Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination*, 113 NW. L. REV. 1163, 1181 (2019).

[207] *Id.* at 1182.

[208] *Id.* at 1183.

[209] *Id.* at 1184.

[210] *Id.* at 1183.

[211] Chandu Siva, *Machine Learning and Pattern Recognition*, DZONE (Nov. 30, 2018) https://dzone.com/articles/machine-learning-and-pattern-recognition.

[212] *Id.*

[213] VICTOR MAYER-SCHONBERGER & KENNETH CUKIER, BIG DATA, A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 53 (2013); Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, 55 COMMS. OF THE ACM 78, 87 (2012).

correlations will provide much more accurate results when compared with a traditional causality-based algorithm, but causality is deeply rooted in human behavior and reasoning patterns.[214] Consider, for example, the case of discovering a correlation between increased ice-cream sales and homicide. There is statistical evidence that demonstrates that when ice-cream sales increase, the homicide rate also increases.[215] Hopefully, such a correlation would not lead police agencies to consider ice-cream consumption in its algorithms, or governmental agencies to decide to limit the ice-cream sales in an effort to combat murder.[216] There could be several reasons for this correlation, ranging from mere coincidence to the warm weather in the summer that both urges people to buy more ice-cream and also makes people angrier and less likely to control their nerves.[217] Thus, an algorithm that considers only factors that cause the outcome reduce the risk of unintentionally including something like ice-cream consumption.

Improved explainability: some researchers claim that relying only on causality would enhance the explainability of the output produced by the algorithm.[218] First, the debate about which factors cause the outcome would have to happen between human who can explain their actions and choices. However, studies show that people consciously and unconsciously weigh more than just legal factors written in a book in their decision making and that their intuitions can often be misleading.[219] Second, the number of factors in an algorithm based causality would naturally be smaller, so situations in which even the engineers who built the algorithm cannot explain which combination of factors led to the end result, let alone policymakers further removed from the algorithm, would be reduced.[220]

Better compliance with constitutional due process requirements: giving that causality based notions enhance explainability and reduce the reliance on solely correlations, they could adhere to due process requirements.[221] Due process protects the individual against the arbitrary use of government power by ensuring that the same process applies to all regardless of race, gender, socioeconomic status, and so on.[222] Due process requires, at a minimum, the right to receive notice, the right to be heard, and the opportunity to respond in a

[214] MAYER-SCHONBERGER & CUKIER, *supra* note 213, at 53.

[215] Justin Peters, *Warm Weather Homicide Rates: When Ice Cream Sales Rise, Homicides Rise. Coincidence?*, SLATE (Jul. 9, 2013, 2:59 PM), https://slate.com/news-and-politics/2013/07/warm-weather-homicide-rates-when-ice-cream-sales-rise-homicides-rise-coincidence.html.

[216] *Id.*

[217] David Trafimow, *The Probability of Simple Versus Complex Causal Model in Causal Analyses*, 49 BEHAV. RES. 739, 743–44 (2017).

[218] *See id.* (describing how causality increases the explainability and gives the models more credibility).

[219] Jennifer Doleac, *Let Computers Be the Judge*, MEDIUM (Apr. 20, 2017), https://medium.com/@jenniferdoleac/let-computers-be-the-judge-b9730f94f8c8.

[220] Zachary C. Lipton, *The Mythos of Model Interpretability*, 16 ACM QUEUE MAG. 30, 35 (2018); *see* Richard A. Berk & Jordan M. Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 FED. SENT'G REP. 222, 222–23 (2015) (explaining the limitations of an algorithm based on causality).

[221] William N. Eskridge, Jr., *Destabilizing Due Process and Evolutive Equal Protection*, 47 UCLA L. REV. 1183 (2000).

[222] *Id*. at 1190.

legal trial.[223] One goal of due process is to strengthen public confidence in the system, which will increase compliance with the law, and if the operation of algorithms is clearer and more coherent, it would increase the likelihood that they would be adopted by the public more smoothly.[224] As a procedural requirement, the judge, or the administrative body taking any action or making a decision in a specific case, has to provide an explanation that consists of two parts: the general rule of law that guided the decision and a specific link and application to the facts of the case.[225] This allows the individual to challenge the decision on one or both grounds. The perception of a fair trial has significant importance to the individual. Thus, people can perceive a trial as fair even if, at the end of the day, they lost—so long as they had a real opportunity to be heard.[226]

The level of explainability, as well as the procedural requirements and the tie to causality varies a lot between different algorithms, and for each social problem, the suitable type of explanation is different.[227] On one end of the spectrum, there will be unexplainable decisions that it will not always be possible to pinpoint the relationship between their factors and the outcome, but still they will be accepted.[228] In the same manner that very important real-life decisions are being determined by a lottery system, such as an army draft or green card determination, there will be decisions that AI algorithms make that do not pose substantial harm to due process, which we will accept. Next, there will be decisions that might not comply with some of the due process requirements, but the benefits of automatizing them using AI will be substantial, so we might consider reconfiguring the balance between due process and other rights.[229] On the other end of the spectrum, there will be decisions suggesting that lack of causality could significantly harm due process, so for them only explainable algorithms will be accepted.[230] Interestingly, when courts were faced with claims regarding violation of due process, they concluded that the use of a risk assessment tool in the sentencing phase of a criminal trial, did not violate the defendant's right to due process so long that the judge use the output of the algorithm as one factor among many others and has the discretion to diverge from its conclusion if needed.[231] This was the outcome of State v. Loomis,[232] a decision from the Supreme Court of Wisconsin. Loomis's claims related to due process centered around the fact that he was sentenced to six years

[223] William Wagner & Joshua R. Castillo, *Friending Due Process: Facebook as a Fair Method of Alternative Service*, 19 WIDENER L. REV. 259, 262 (2013).

[224] Tracey L. Meares, *What's Wrong with Gideon?*, 70 U. CHI. L. REV. 215, 216–17 (2003).

[225] Franck I. Michelman, *Formal and Associational Aims in Procedural Due Process*, 18 NOMOS 126, 132–33 (1977).

[226] Josh Bowers & Paul H. Robinson, *Perceptions of Fairness and Justice: The Shared Aims and Occasional Conflicts of Legitimacy and Moral Credibility*, 47 WAKE FOREST L. REV. 211, 211–12 (2012).

[227] Trafimow, *supra* note 217, at 743.

[228] *Id.* at 744.

[229] Eskridge, *supra* note 221, at 1201.

[230] *Id.*

[231] Recent Case, *Criminal Law—Sentencing Guidelines—Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing—State v. Loomis, 881 N.W.2d 749* (Wis. 2016), 130 HARV. L. REV. 1530, 1532–33 (2017).

[232] State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

in prison based on an assessment of a proprietary algorithm that its methodology in calculating the risk is a trade secret.[233] The court concluded that since the information that the algorithm relies on in the assessment are publicly available and data from a questionnaire filled by Loomis himself, so he could verify the accuracy of the information even though the exact way the algorithm calculated the score is not known. Nevertheless, the court warned that these algorithms should be used cautiously, the algorithm cannot be the determinative factor in deciding if the defendant will be incarcerated or the severity of the sentence, and the algorithm should be validated for the population it is used on.[234] It is important to mention that there was an attempt to challenge this decision in the U.S. Supreme Court, but certiorari was not granted.[235] Hence, the court set a high bar on violation of due process by algorithms.[236] Interestingly, the instructions that the court set are directed more toward the judges as the users of the algorithm, but not much on the design of the algorithm and the type of factors that could be included, if they cause the outcome; or the way to group them. As it will be explained below, addressing algorithmic fairness is a joint task for both developers and policymaker, and every actor in the field has a rule in implementing some needed changes.

B. The Challenges of This Approach

Despite the fact that causal reasoning-based approaches seem appealing, they are not easy to implement.

First, the causal diagrams that feed the model are very sensitive; they require identifying the important variables for the matter and taking a normative stance about causality between variables and between each variable and the outcome.[237] This is a complex task that only experts in the particular domain are capable of.[238] To try to leave this task in the hands of computer scientists and statisticians would be very dangerous because any determination about causality will have a direct impact on the outcome, and causality is not easy to detect.[239] The first step of drawing the causal diagram has significant importance because counterfactual fairness does not guarantee causality; it assumes that the causal diagram is correct and then assesses what would have happened if one factor is replaced with a counterfactual one.[240]

Second, it is not easy to conclude anything about the real world based on the counterfactual world.[241] The interaction between the different factors in the real world is too complex, and changing just one of the factors in the

[233] *Id.* at 761.

[234] *Id.* at 769.

[235] State v. Loomis, 881 N.W. 2d 749 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

[236] Anne L. Washington, *How to Argue With an Algorithm: Lessons from the COMPAS-ProPublica Debate*, 17 COLO. TECH. L. REV. 131, 132–34 (2019).

[237] Kohler-Housman, *supra* note 206, at 1177.

[238] *Id.* at 1221.

[239] *Id.*

[240] *Id.* at 1211.

[241] *Id.*

counterfactual world will not necessarily lead to fairer results in the real world.[242]

Third, according to counterfactual causality, race can be manipulated only if the manipulation does not change other aspects of the life of the individual, however . . . the term 'race' cannot refer to an attribute, a genetically produced trait, or a signifier—level of melanin in skin, phenotype, distinctive names or speech—that people just have and thereby obviously belong to a designated racial group. The term references a complexly constituted social fact, whereby material and dignitary opportunities are organized such that certain physical and cultural signifiers become the salient markers of consequential cultural categories, and those categories are constituted by a constellation of social relations and meanings with a definite content and organization.[243]

Fourth, discrimination is a "thick ethical concept" that requires "complex social knowledge in order to be used and decoded."[244] In order to evaluate if a particular action is discriminatory, we would need to have a comprehensive knowledge about institutional and cultural facts, and take a normative stance about the way those institutions operate.[245] We would need to have sociological and anthropological knowledge about race and its implications, and about what is considered fair and just in society.[246]

Fifth, counterfactual studies that audit the racial component do not measure discrimination properly because different people may value race differently and associate it with different qualifications.[247] Imagine, for example, that an identical job application was sent to different employers: One applicant has a white-sounding name and the other a black-sounding name. One employer may value the achievements of the candidate regardless of race, and appreciate the fact that he built a successful career despite growing up in a low-income neighborhood and going to a very low-rated high school. Another employer might look down at the candidate based on the connotations associated with his name and see the achievements as merely the expensive cost of affirmative action. Therefore, it is very hard to measure the effect of race.

V. Interim Summary

Before turning to the part that discusses the practical implications of choosing one fairness notion over the other for both policymakers and developers, the table below summarizes the fairness notions discussed above as well as their applicable legal mechanism; and an example of a policy domain that they could be suitable for. It is important to mention that each one of the fairness notions can be the subject of a whole paper; and that the policy domains were discussed only on a high level of generality. When conducting a deeper

[242] *Id.* at 1205.
[243] Kohler-Housman, *supra* note 206, at 1170.
[244] *Id.* at 1221–22.
[245] *Id.* at 1171.
[246] *Id.* at 1176–77.
[247] *Id.* at 1211–12.

analysis of a certain policy domain, attention to specifics would matter for the chosen notion of fairness. For example, in the context of criminal justice, it is highly plausible that a different notion of fairness will be suitable for different stages, such as pretrial, sentencing, and parole. But the goal of the typology presented in this paper is to provide a flexible framework for the discussion about the matching between the notion of fairness, the legal mechanism and the policy domain at stake.

Table 2: Notions of fairness and summary of their corresponding legal mechanisms.

Notion	**Sub-notion**	**Corresponding Legal Mechanism**	**Example of an Implementable Case**
Individual fairness	The unaware approaches	Equal opportunity as colorblindness	When the population is homogeneous like in marketing a luxury product
	Fairness through awareness	Equal opportunity based on similarities, and levels of scrutiny	A variety of high-stakes cases like criminal justice and credit scoring but only if the metric is agreed upon
Group fairness	Decoupling	Affirmative action (as separate but equal)	When the minority group is very small and has unique characteristics like women in the criminal justice system
	Statistical or conditional parity	Affirmative action (preferably through critical diversity)	Cases where affirmative action was approved by law like in hiring and school admissions
	Equal opportunity	Affirmative action (via equality of opportunity)	When fixing the outcome is sufficient and does not require fixing the process that led to this outcome
	Equalized odds	Achieving equality by equalizing false positive and false negative errors	When it is possible to achieve the right balance between the two types of errors
	Calibration	Achieving equality by statistical significance	High-stakes cases that society is willing to give up on equalizing the error rates
	Multicalibration	Achieving equality by statistical significance, and accounting for intersectionality	Pretrial, but it should be applied cautiously since it is a new notion
Causal reasoning	Counterfactual fairness	Disparate treatment and disparate impact analysis	Where the counterfactual factors are traceable and do not significantly impact other factors

VI. BALANCING BETWEEN DIFFERENT NOTIONS OF FAIRNESS

It should not come as a surprise that, from both a policy and technical perspective, satisfying several notions of fairness simultaneously is mutually incompatible.[248] In the same manner that it is not possible to maximize profit for borrowers and open new doors for historically disadvantaged minorities, or to keep every defendant behind bars in the name of public safety and respect for individual justice principles, it is also not possible to satisfy statistical parity and fairness through awareness at the same time, or to satisfy equalized odds and calibration simultaneously. Choosing one notion of fairness over the other has a direct impact on the case of the individual as well as far-reaching policy implications.

The goal of this section is to go back to the COMPAS example mentioned in the introduction and use pretrial as a case study for illustrating the implications of applying each one of the notions of fairness discussed in this paper in this domain.

A. The COMPAS Example

COMPAS is an empirical risk and needs assessment tool integrated in the Northpointe Suite, a web-based assessment and case management system for criminal justice practitioners.[249] COMPAS is being used by criminal justice agencies in different stages of criminal justice such as pretrial, sentencing, jail placement, probation, and parole.[250] Based on a combination of dynamic and static factors for each defendant, COMPAS provides a risk score on a scale of low to high, and this score represents the likelihood that the defendant will recidivate.[251] COMPAS was developed by the for-profit company Northpointe, today doing business as Equivant.[252] The software operates as a black box, meaning that the inner workings of the algorithm, as well as the way the score is calculated, are not known to the public or to defendants.[253] The news outlet ProPublica obtained the risk scores assigned by COMPAS to more than 7,000 people arrested in Broward County, Florida and examined the fairness of COMPAS classifications.[254] ProPublica looked at how many defendants were charged with new offenses after two years of their release.[255] ProPublica concluded that COMPAS is biased because the false positive rate was much higher among black defendants.[256] The algorithm falsely labeled black defendants as future criminals twice as much as it did so for white defendants.[257]

[248] Berk et al., *supra* note 10, at 1; Chouldechova, *supra* note 10, at 153–54; Friedler et al., *supra* note 10, at 329; Kleinberg et al., *supra* note 10, at 40; Narayanan, *supra* note 10.

[249] EQUIVANT, *supra* note 1, at 1.

[250] *Id.*

[251] *Id.* at 1–2.

[252] *Id.* at 2.

[253] Angwin et al., *supra* note 2.

[254] *Id.*

[255] *Id.*

[256] *Id.*

[257] *Id.*

Among black defendants, 42% of those who were released from jail and did not commit any future crimes were wrongly labeled high-risk, while among white defendants the algorithm made the same mistake in only 22% of cases.[258] In other words, examining the accuracy of COMPAS's classifications after two years revealed that the algorithm made classification errors among both blacks and whites and unnecessarily labeled defendants as high-risk even if they were not. However, the mistakes occurred for black defendants twice as much as compared to white defendants.[259] ProPublica also identified false negative errors, meaning that COMPAS falsely flagged white defendants as low-risk (although the exact percentage was not mentioned in the article).[260]

Northpointe dissented from the findings, and published their own investigation showing how COMPAS is equally fair to black and white defendants.[261] The rebuttal attracted the attention of many academics and other news outlets.[262]

In the center of the debate was a disagreement about fairness definitions. ProPublica claimed that COMPAS is biased because ProPublica believes in the fairness definition of equal opportunity.[263] They perceived as unfair that low-risk black defendants did not get an equal opportunity to be labeled as such.[264] According to ProPublica, it is not fair that COMPAS makes this serious error more frequently for one race group than for another.[265] To the contrary, Northpointe claimed that their algorithm is calibrated, meaning that each one of the scores from 1–10 given by COMPAS actually reflects equality.[266] For example, among defendants who were scored a 7, 60% of white defendants and 61% of black defendants recidivated.[267] In addition, Northpointe claimed that COMPAS is fair since, as directed by the law, it does not take into account race explicitly.[268] Lastly, Northpointe pointed out that, given the different base rate of black and white defendants, the gap that ProPublica referred to will always exist regardless of COMPAS.[269]

As explained above, it is not possible to satisfy both notions of fairness simultaneously.[270] Neither ProPublica nor Northpointe are addressing the problem of the different base rate among black and white defendants. If ProPublica is vouching for equal opportunity, that still does not guarantee equality to high-risk defendants who might also be wrongly classified; if the results are not calibrated this might lead judges to treat black and white

[258] *Id.*

[259] Corbett-Davies, *supra* note 4.

[260] Angwin et al., *supra* note 2.

[261] DIETERICH ET AL., *supra* note 5, at 2–4.

[262] Corbett-Davies, *supra* note 4; Flores et al., *supra* note 6, at 39–40.

[263] Corbett-Davies, *supra* note 4.

[264] *Id.*

[265] *Id.*

[266] *Id.*

[267] *Id.*

[268] *Id.*

[269] *Id.*

[270] Friedler et al., *supra* note 10, at 329; Berk et al., *supra* note 10, at 1; Kleinberg et al., *supra* note 10, at 40; Chouldechova, *supra* note 10, at 153–54; Narayanan, *supra* note 10.

defendants differently and to not trust the algorithm. As for Northpointe, their defense is centered on the legal requirement to not explicitly consider race, on technical errors made by ProPublica, and on the general statistical validity of all score levels.[271] Northpointe derives the COMPAS score from a questionnaire that includes more than 130 questions and goes into in-depth details about the defendant's life and family.[272] It is highly feasible that if the questions were tuned to be more inclusive and race sensitive, the results would be fairer; since unfairness in machine learning could occur not only because of the way the algorithm is designed but also because of the way data is being collected.

What Would Have Been the Outcome if COMPAS Used Different Fairness Notions?

It is hard to know exactly what would have resulted if Northpointe chose to implement a different definition of fairness in COMPAS. Below is an estimation of the potential practical impact of each fairness notion discussed in this paper.

The unaware approach—COMPAS does not take race into account given the legal prohibition on doing so, but that does not mean that COMPAS is satisfying individual fairness.[273] The questionnaire used to generate COMPAS's score reveal many factors that can be a proxy for race. For example, the questionnaire asks defendants if one of their family members or friends have ever been sent to jail or prison.[274] Given the over-representation of black defendants in the criminal justice system and the vicious circle of poverty among black communities, it is plausible to assume that the answer to such a question and many others will be a proxy for race. Removing all the questions that can serve as a proxy for race from the questionnaire is nearly impossible since, as mentioned earlier, race is a complex social fact that affects all life experiences and even questions that on their face do not seem associated with race can be a proxy.[275] Nevertheless, even if all proxies can be removed, it is not possible that in bail determination the groups of individuals that we are comparing amongst will be homogenous. The severity of the crime on its own is a very dynamic factor that requires treating different groups of individuals differently.

Fairness through awareness—This could be an interesting case study that might enhance fairness. ProPublica found that only 20% of people predicted to commit violent crimes went on to do so.[276] The idea of having the severity of the crime as a criterion for classification could be factored in the metric. It would be interesting to compare defendants who commit similar crimes or have similar priors and observe whether the results of a metric taking these factors into consideration would be fairer. Other factors that the metric might flag could be age or gender. One option is to frame the discussion about the metric around

[271] DIETERICH ET AL., *supra* note 5, at 23–31,

[272] Angwin et al., *supra* note 2.

[273] Corbett-Davies, *supra* note 4.

[274] Angwin et al., *supra* note 2.

[275] Kohler-Housman, *supra* note 206, at 1170.

[276] Angwin et al., *supra* note 2.

the levels of scrutiny and to say for example that the severity of the crime could always be included but criteria like gender and race will be subject to approval by a higher authority at the criminal justice agency and the court should be notified about the use. In any case, this will spark a heated debate about the differences and similarities between individuals in order to determine the metric. Such debate needs to take place publicly among experts in the field and not behind the closed doors of Northpointe. The metric has to be open and public, even though this goes against the business model of Northpointe, which benefits from a proprietary algorithm.[277] A middle ground solution could be to make the metric public but allow the inner workings of the algorithm to remain proprietary.

Decoupling—In pretrial, decoupling would mean that just as there is COMPAS Women, there will be COMPAS Black and COMPAS White. The question is whether we would allow this as a society. First, given the ruling in *Loomis* that classification on the basis of gender does not constitute violation of due process, it is not clear if racial classification would be viewed in the same way.[278] Second, although classification on the basis of race could seem like an explicit discrimination, as described above, not every use of race constitutes disparate treatment.[279] Racial discrimination permeates the criminal justice system at all stages. Black Americans are getting arrested three times more than would be expected based on their percentage of the population; in plea bargaining, black defendants achieve fewer reductions in their sentences compared to white defendants, convicted black defendants are more likely to receive longer sentences, and black Americans are much more likely to experience police brutality.[280] Decoupling aims to create a separate mechanism for providing a more fair recidivism score for black defendants. Nevertheless, flagging black defendants as a troubled community that require separate treatment could easily broaden the gap between the majority group and minorities and will not yield a positive outcome, neither rhetorically nor practically. In addition, intersectionality will make it hard to determine how to classify defendants and which algorithm to apply to their case if they belong to more than one minority group or if they do not associate themselves with the group they are classified under. The example of jury verdicts illustrates some of the negative consequences of differential treatment. Research show that white jurors are not the only ones more likely to find a black defendant guilty compare to a white defendant—even though jurors are tempted to be unjustifiably more lenient when both the defendant and the jurors are of the same race, and differences due to the victim's race have also been reported.[281] In other words, differential racial treatment can also cause divergence from the rule of law.

[277] Corbett-Davies, *supra* note 4.

[278] *See generally* Jason R. Bent, *Is Affirmative Action Legal?*, 108 GEO. L.J., at 55 (forthcoming 2020) (discussing the legality of algorithmic based affirmative action); Angwin et al., *supra* note 2.

[279] *See supra* Section III.C.

[280] *The Color of Justice*, CONST. RIGHTS FOUND., https://www.crf-usa.org/brown-v-board-50th-anniversary/the-color-of-justice.html.

[281] Norbert L. Kerr et al., *Defendant-Juror Similarity and Mock Juror Judgments*, 19 LAW & HUM. BEHAV. 545, 545–46, 555 (1995).

Statistical parity and conditional statistical parity—Aiming for affirmative action in the context of criminal justice is problematic. As described above, at the core of a fair law enforcement system is individual justice.[282] Being more lenient toward black defendants at the "expense" of white defendants means technically having different thresholds for release for black and white defendants, an outcome that goes against basic principles of equal justice.[283] It is not possible to decide for example, that in order to have an equal number of black and white defendants released in pretrial, black defendants that are ranked by COMPAS below level 6 will be released while among white defendants only those ranked below 8 will be released. In other words, since the base rate of those groups is not equal, statistical parity is hard to achieve without harming the other community, and in the context of COMPAS, harming means unnecessarily putting someone in jail or releasing a risky defendant.

Equal opportunity—This notion ensures that only defendants that are low-risk will have an equal chance to be classified as a low risk. The most vulnerable segment of defendants is comprised of those who will be wrongly classified as high risk due to the negative ramifications of spending unnecessary time in jail. Thus, practically speaking, this notion is undesirable in the context of pretrial, since it does not address the needs of high-risk defendants to be classified correctly. In the criminal justice system, favoring false positives over false negatives or vice versa is not plausible because both are equally important.[284]

Equalized odds—This is the notion that ProPublica was arguing for, equalizing the error rates that COMPAS makes among black and white defendants. As mentioned before, despite its appeal, applying this notion in pretrial is not an easy task since policymakers will have to quantify the line between public safety and individualized justice. What is the threshold that our society is willing to tolerate? This is a decision that policymakers and judges, rather than Northpointe, should make. In an interesting experiment, researchers attempted to implement the Blackstone ratio—better ten guilty men go free than one innocent person suffer, and equalize it for both blacks and whites.[285] The result is that yet again, we end up with different thresholds for the two groups, while white defendants get jailed for a risk score of 7, black defendants who are scored 7 are being released.[286] To that we should add the fact that Blackstone ratio on its own compromises public safety significantly.[287] Currently, for black defendants the error rate of the algorithm is 42% and for white defendant the error rate is 22%.[288] If we were to equalize the error rates, it would have to be

[282] *See supra* Section VI.C.

[283] *See* MacCarthy, *supra* note 11, at 115–22 (detailing an overview of philosophical theories relating to distributive justice).

[284] *See* Bowers, *supra* note 226, at 247–50 (discussing the problems associated with false positives and false negatives).

[285] Karen Hao & Jonathan Stray, *Can You Make AI Fairer Than a Judge? Play Our Courtroom Algorithm Game*, MIT TECH. REV. (Oct. 17, 2019), https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

[286] *Id*.

[287] *Id*.

[288] Corbett-Davies, *supra* note 4.

at the expense of calibration. If the algorithm is not calibrated and suppose it provides a more accurate predictions for white defendants, judges will learn not to trust the algorithm when it comes to predictions about black defendants.

Calibration—This is the notion that Northpointe is actually satisfying, making sure that in each score level the percentage of black and white defendants that are recidivating is equal.[289] These examples show that calibration is important but not sufficient.[290] Since the base rate is different, and black defendants have a higher recidivism rate on average, calibration is not sufficient to prove that the algorithm is not biased; and this was causing the difference in the error rate.[291]

Multicalibration—This could be a notion that balances properly between individual and group fairness. It will make sure that all risk levels are calibrated, and by using decoupling and creating slightly different thresholds for different groups, it will attempt to ensure that intersectionality is being respected and each individual's special characteristics are calibrated.

Causal reasoning—In order to satisfy this notion, Northpointe has to be very explicit about the factors that are taken into account in COMPAS and also explain what purpose each question in the long questionnaire is serving. While this might add some clarity to the overall outcome and might spark a debate about the necessity of certain questions, this will also not be sufficient because, as described above, it is hard to measure causality and to track down the counterfactual impact of significant factors such as race and gender.

In conclusion, the example of COMPAS illustrates that all notions of fairness come at a price. The disparities draw attention to where bias might exist and force us to have a conversation about what kind of tradeoffs we are willing to make. The fact that multicalibration might be the most appropriate solution in this particular case does not imply that other notions of fairness are not valid. For each policy domain, it is important to do this type of exercise and assess what the most suitable notion of fairness is given the legal and social framework. In addition, criminal law enforcement is very decentralized in the United States and policies and regulations are different across jurisdictions; and those differences could impact the chosen notion of fairness.[292] For example, unlike most jurisdictions where in the pretrial stage judges are required to take into account the risk that the defendant will commit another crime and the risk that they will not attend their trial, in New York the only factor that judges are allowed to consider is the risk of failure to appear for the trial.[293] Therefore, the notion of equalized odds for example will be easier to implement there since it is clear what should be favored when balancing public safety and individual justice. This exercise of finding the right notion of fairness should include the

[289] *Id*.

[290] Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV 1808.00023 1, 16–17 (2018).

[291] *Id*.

[292] *See* Angwin, *supra* note 2 (discussing the different ways risk assessment is used throughout the United States).

[293] Kleinberg et al., *supra* note 183, at 246.

relevant regulation and case law of each jurisdiction that intend to implement a fair algorithm.

B. Implications for Developers

1. Clarifying Their Approach to Fairness

It was not until ProPublica published its story that Northpointe released its technical paper, which shed some light on the validity of COMPAS.[294] Nevertheless, Northpointe does not explicitly mention their notion of fairness, and most of the paper is devoted to technical mistakes made by ProPublica.[295] Being explicit about the notion of fairness implemented in the algorithm, or even better, detailing if other notions were tested before the current one was chosen and what the impact is of each notion on different segments of society, will help all actors in the field in determining its applicability.[296] Many validation techniques have been developed in computer science literature in order to minimize the gap between accuracy and fairness.[297] Such methods are meant to deal with some of the causes of unfairness in the algorithms, for example, the fact that the training data does not represent the whole population.[298] It is very important that techniques for detecting and avoiding bias be implemented in all the validation processes, starting from the testing phase to the formal verification phase.[299] If possible, developers should publicly communicate the precautionary measurements related to fairness that are taken in each phase of the development of the algorithm. To date, validation of the machine learning algorithm often refers to the level of accuracy that the model performs—the error rate of the machine learning model—but an accurate model is not necessarily a fair model.[300]

Another approach that big tech companies are taking is developing meta tools for enhancing fairness, preventing, and mitigating bias.[301] For example, IBM offers a service that runs on the IBM cloud, which provides customers with insights about the decision-making process of the algorithm; it explains which factors influence the decision and assesses the level of accuracy and fairness of

[294] DIETERICH ET AL., *supra* note 5.

[295] *See id.* at 13–14 (discussing mistakes made by the ProPublica article concerning COMPAS risk scales and its impact on accuracy equity).

[296] Nicholas Diakopoulos & Sorelle Friedler, *How to Hold Algorithms Accountable*, MIT TECH. REV. (Nov. 17, 2016), https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/.

[297] Ajitesh Kumar, *Machine Learning: Validation Techniques*, DZONE (Feb. 12, 2018), https://dzone.com/articles/machine-learning-validation-techniques.

[298] *Id*.

[299] Ian Goodfellow & Nicolas Papernot, *The Challenge of Verification and Testing of Machine Learning*, CLEVERHANS-BLOG (June 14, 2017), http://www.cleverhans.io/security/privacy/ml/2017/06/14/verification.html.

[300] Aaron Roth, Institute for Advanced Study, *Quantifying Tradeoffs Between Fairness and Accuracy in Online Learning: Computer Science/Discrete Mathematics Seminar I*, YOUTUBE (Jan. 30, 2017), https://www.youtube.com/watch?v=tBpd4Ix4BYM.

[301] Kyle Wiggers, *IBM Announces Cloud Service to Help Businesses Detect and Mitigate AI Bias*, VENTUREBEAT (Sept. 19, 2018), https://venturebeat.com/2018/09/19/ibm-announces-cloud-service-to-help-businesses-detect-and-mitigate-ai-bias/.

the algorithm.[302] Facebook and Microsoft offer similar services that aim to automatically detect bias and alert data scientists to any potential biased behavior of the algorithm.[303] In addition, there are many educational sources that are meant to assist developers in detecting biases.[304] While these tools could be useful and might be able to point out potential problematic behavior of algorithms, they cannot be used alone, and should be taken with a grain of salt because mitigating bias cannot be fixed by a miracle. Every algorithm that is implemented to solve a social problem is operating in a complicated environment and affected by many different aspects. The way to detect bias in the medical sphere is probably not the same as detecting bias in the criminal justice system, and the implications of fairness in each context are different. Any bias-detecting tool lacks the complete context and broader picture, so it will not be able to detect certain biases. For example, Neil and Winship show how the methods used to determine whether the police discriminate on the basis of race do not produce valid inferences and they may indicate discrimination when it is not present or vice versa.[305] The most commonly used tests for discrimination are the benchmarking test and the outcome test.[306] According to the benchmarking test, we should compare the frequency in which different races experience discriminatory police contact such as stops, searches, or arrests.[307] The outcome test measures the rates in which contraband is found across races, and if contraband is found among one race at a significantly low rate, this implies that their search threshold is lower.[308] The problem with those tests is that they have the potential of simplifying the assumptions made about police behavior, and such simplification could undermine the validity of the conclusion about discrimination.[309] In the context of police behavior, stops are a function of who is on the street, at what time of day, where the search is taking place, and which crimes are the target of the police. If we are focusing on stop and frisk of pedestrians, it is safe to assume that the poor often cannot afford owning a car, so they will walk more and be exposed to more stops compared to the wealthy, who will not be within the radar. Similarly, if the police focus on high-crime neighborhoods where the distribution of the population is not equal, then statistically, more black people will be arrested.[310] Of course, there are discriminatory reasons that lead certain groups of the population to be poor, to live in high-crime neighborhoods, or to be unable to afford a car; these historical and systemic reasons will affect the outcome.[311] The study by Neil and Winship

[302] *Id*.

[303] Dan Robinski, *Microsoft Announces Tool To Catch Biased AI Because We Keep Making Biased AI*, FUTURISM (May 25, 2018), https://futurism.com/microsoft-announces-tool-catch-biased-ai.

[304] Roland Neil & Christopher Winship, *Methodological Challenges and Opportunities in Testing for Racial Discrimination in Policing*, 2 ANN. REV. CRIMINOLOGY 73, 157 (2019).

[305] *Id.* at 74–75.

[306] *Id*.

[307] *Id*.

[308] *Id*.

[309] *Id.* at 308.

[310] *Id.* at 78, 308.

[311] *Id.* at 78.

illustrates nicely the limitations of a very well-used statistical test.[312] The social problems that we are deploying algorithms in have many layers of complexity that a simple statistical test will not be able to detect. This is not to say that we should give up on the benchmark or the outcome tests, rather we should learn how to exploit their strengths and be aware of their vulnerabilities. If the algorithm shows that the police stop and search disproportionately more blacks than whites, we should dig deeper into the problem and try to understand the reasons as well as the findings.

One way to compile all the different tools for mitigating bias and augmenting fairness is by creating a questionnaire or impact assessment tool, similar to the Privacy Impact Assessment process that companies has to publish if their products entail risk to the right for privacy. In fact, the Canadian government is in the process of developing and implementing an Algorithmic Impact Assessment tool [AIA].[313] The tool aims to "help institutions better understand and mitigate the risks associated with Automated Decision-Making Systems by providing the appropriate governance, oversight, and audit requirements."[314] The tool consists of a questionnaire of 57 questions about the business process, system design and data maintenance.[315] It takes approximately 35 minutes to complete and designers are allowed to fill it as many times as needed throughout the design process in order to make sure that they end up with the best product.[316] Upon completion, the tool scores the algorithm on a scale of 1–4, and the higher the score there will be more requirements on the developers to fulfil before the algorithm can be used.[317] The tool measures the impact of the algorithm on legal and ethical issues including procedural fairness and the impact on individuals as well as groups.[318] The tool has been improved constantly and its implementation is in its early phase, so it is not possible yet to assess its effectiveness but the government of Canada has been soliciting interdisciplinary feedback in order to make it as useful as possible for all.[319] As the tool will be more developed, more questions related to fairness can be added and this could be an interesting way to demand more transparency from developers about the chosen notion of fairness that they implemented in the algorithm and its applicability to equal protection and due process. A similar regulatory effort, although much narrower in its scope, is making its way in the U.S. Congress.[320] The Algorithmic Accountability Act, introduced in April

[312] *Id.* at 159.

[313] Michael Karlin & Noel Corriveau, *The Government of Canada's Algorithmic Impact Assessment: Take Two*, MEDIUM (Aug. 7, 2018), https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f.

[314] *Id*.

[315] ALGORITHMIC IMPACT ASSESSMENT (ARCHIVED), GOVERNMENT OF CANADA DIGITAL PLAYBOOK (DRAFT), GOVERNMENT OF CANADA, https://canada-ca.github.io/digital-playbook-guide-numerique/views-vues/automated-decision-automatise/en/algorithmic-impact-assessment.html (last visited Nov. 29, 2019).

[316] *Id*.

[317] Karlin & Corriveau, *supra* note 313.

[318] *Id*.

[319] *Id*.

[320] Joshua New, *How to Fix the Algorithmic Accountability Act*, CENTER FOR DATA INNOVATION, (Sept. 23, 2019), https://www.datainnovation.org/2019/09/how-to-fix-the-algorithmic-accountability-act/.

2019, would authorize the FTC to develop regulations that require companies to conduct impact assessment if they create algorithms that pose high risk automated decision systems.[321] The act is meant to address some of the concerns that algorithmic decision making are raising, and to mitigate bias and discriminatory impact.[322] However, unlike the Canadian regulation, there is no guidance on the type of the impact assessment that should be conducted, and there is no one questionnaire that all companies will be subjected to. In addition, the focus here is only on big companies that pose high-risk and the companies will not be required to disclose the assessment.[323]

2. Increased Social and Cultural Understandings

According to some research, without prior exposure to different philosophical approaches, most people will be utilitarianists by default because of the tendency to link utilitarianism with wealth production.[324] But when looking closer at ethical dilemmas arising in the interaction between humans and machines, different geographical-based preferences can be observed.[325] A group of researchers from the MIT Media Lab surveyed more than two million participants from more than 200 countries, presenting them with different variations of the trolley problem, where a driverless car has to choose between two fatal options.[326] The study pointed to some consensus in regard to public principles of favoring human life over the life of an animal, and choosing to preserve the lives of more people rather than a few.[327] However, the study also highlighted significant differences between countries and geographical regions.[328] For instance, in the eastern cluster of countries (including Asian countries), people preferred to preserve the life of the elderly, in comparison with the western cluster of countries where there was a clear tendency to favor younger people.[329] In addition, countries belonging to the southern cluster (Latin American countries and countries with French influence) preferred to preserve females' lives over males.'[330]

The study demonstrates how moral preferences differ across cultures, and how certain norms that seem obvious for one culture are valued less by others.[331] People in individualistic cultures might value privacy more than people in collectivistic cultures, and they might favor individual fairness notions over group fairness notions. These things should be taken into account by the

[321] *Id*.
[322] *Id*.
[323] *Id*.
[324] Emanuelle Burton et al., *Ethical Considerations in Artificial Intelligence Courses*, 38 AI MAG. 22, 26–28 (2017).
[325] Edmond Awad et al., *The Moral Machine Experiment,* 563 NATURE 59, 61 (2018).
[326] *Id*. at 59.
[327] *Id.* at 60.
[328] *Id.* at 62.
[329] *Id.*
[330] *Id.*
[331] *Id.*

algorithm.[332] The values which societies and their members subscribe to are different, and what is deemed moral or ethical in one society might be considered the total opposite in another. Although equal opportunity and due process are recognized as important values in all democratic societies, their implications might be different. This is important, as the legal and social standing point in regard to a particular case will determine the fairness-correcting method that we will choose.[333] Each one of the legal mechanisms discussed in this paper is designed to achieve a certain policy goal. For example, as it can be inferred from the discussion about group fairness above, the common denominator of notions in this group is their connection with affirmative action, a mechanism for improving the position of historically disadvantaged minority groups in society. As the hiring example in statistical parity demonstrate, most developers who lack legal knowledge would understand affirmative action in a narrow way, as a mechanism that requires quotas and urge employers to hire unqualified women. However, as explained, there are many different ways of achieving affirmative action. Understanding the legal and social circumstances that the algorithm will operate in would tilt a better balance between short- and long-term policy goals.

Grasping a basic understanding of different philosophical theories will broaden the toolkit of computer scientists and will ensure that the algorithms that we increasingly implement in our daily life are better aligned with our social and cultural norms. But this too has to be done cautiously as philosophers such as Aristotle and Rawls did not build theories for understanding fairness or justice at a level as narrow as a particular algorithm.[334] Philosophical theories usually refer to the structure of the social, political, and cultural systems, and applying a general theory to a very particular case could be tricky.[335] However, familiarizing developers with different theories of philosophy and applying them to the particular domain of the algorithm could be beneficial.

In light of recent scandals involving unethical use of technology, there is a growing recognition of the need to incorporate more ethical, legal, and social understandings into the curricula of computer science programs and job training for engineers; and we can only hope that those initiatives will continue to grow.[336] There has been a traditional separation between disciplines considered to heavily implicate ethics and therefore demand ethical training, such as the job of philosophers, and fields like computer science, traditionally viewed as needing training for ethics-devoid tasks like building models and making sure

[332] Jess Whittlestone et al., *The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions*, AIES ‘19: PROCEEDINGS OF THE 2019 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 195, 197 (2019).

[333] Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, 81 PROCS. MACHINE LEARNING R., CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1, 1–2 (2018).

[334] *See id.* at 2 (noting that “it is therefore unsurprising that attempts to formalise fairness in machine learning contain echoes of these old philosophical debates”).

[335] *See id.* at 3 (describing how “political philosophy” is currently referred to as “inspiration” only “in a limited and somewhat ad-hoc way” in current studies of machine learning and fairness, which suggests that making such connections can be difficult).

[336] Casey Fiesler, *What Our Tech Ethics Crisis Says About the State of Computer Science Education*, BERKMAN KLEIN CENTER (Dec. 5, 2018), https://howwegettonext.com/what-our-tech-ethics-crisis-says-about-the-state-of-computer-science-education-a6a5544e1da6.

that they work properly.[337] In the course of a computer science education, ethics is usually taught as a stand-alone course toward the end of the degree after the students have acquired most of the needed technical knowledge.[338] However, this model of teaching reinforces the idea that ethics is a separate matter for others to worry about.

It is important to bear in mind that philosophy is one tool among many that computer scientists need to have in their toolbox while designing an algorithm. Law, public policy, economics, political science, and others are equally important. Engineers are not expected to become lawyers, policymakers, or economists, but having a full picture of the social and cultural implications of the algorithm is very important.

C. Implications for Policymakers

1. Clarifying the Laws and Policies

There is an inherent difference between the way machine learning-based algorithms operate and the way regulation is designed. Most machine learning algorithms that exist today fall in the category of narrow AI and are task-specific.[339] While they can generalize pattern recognition to some extent, their capabilities are not even close to those of the human mind or what researchers call Artificial General Intelligence, the futuristic type of algorithms that will be capable of sophisticated decision making and reacting to unexpected and complex environments.[340] Narrow AI algorithms perform best when we define, in as many details as possible, the future outcome that the algorithm is expected to predict. Contrarily, laws and policies are designed to be broad and flexible enough in order to be applicable for a wide variety of cases, and to adapt to the changing needs and circumstances of society. Regulation often represents a balance between competing values of different actors in the field.[341] This makes it challenging to give computer scientists clear instructions on how to design algorithms and incorporate decision-making in them.

One of the unique characteristics of regulation and policy is the built-in room for discretion. Laws and policies leave significant room for discretion in order to allow public officials to enhance individual justice and achieve a more just result that matches the facts of the case to the requirements of the law. Discretion allows public officials the power to interpret the existing laws in order to achieve societal goals and to provide solutions to cases where the law is silent.[342] The room for discretion allows public officials to utilize their

[337] *Id.*

[338] *Id.*

[339] Peter Voss, *From Narrow to General AI*, MEDIUM: INTUITION MACHINE (Oct. 4, 2017), https://medium.com/intuitionmachine/from-narrow-to-general-ai-e21b568155b9.

[340] *Id.*

[341] Neil Gunningham & Darren Sinclair, *Smart Regulation*, *in* REGULATORY THEORY: FOUNDATIONS AND APPLICATIONS 133, 133–34 (Peter Drahos ed., 2017).

[342] Charles H. Koch Jr., *Judicial Review of Administrative Discretion*, 54 GEO. WASH. L. REV. 469, 475 (1985).

experience and professional intuition in order to weigh in on and achieve a more just result. However, discretion does not lead to only positive interventions, public officials can also use their discriminatory power to reinforce biases, stereotypes, and prejudices. For example, studies show that when judges rely on their intuitions, they do not use information reliably; they may assign weight to items that are in fact not predictive, or they may be overly influenced by causal attributions.[343] This is not to say that judges are being explicitly discriminatory. The human brain is a black box, and psychology research shows that people who discriminate are often not aware of it because there are rapid automatic responses that the brain generates before conscious can intervene.[344]

Replacing the traditional decision-making process with an algorithm or deploying an algorithm to assist the decision makers raises several concerns.

First, a preliminary discussion about whether each particular decision can be altered with AI should take place, and the answer will not always be yes.[345] The examples discussed in this paper lead to the conclusion that if AI is to be used for making binding determinations, it can only be done where the laws and regulations are detailed and clear, or when we are confident that computing the rules will work well. In cases where there is a lot of room for tacit knowledge retained by experienced decision makers, we cannot replace them with AI; at most we can use AI to assist them. Thus, when the fairness notion that we are following is statistical parity, conditional statistical parity, or any notion that defines a clear-cut number (in the outcome), automating the decision is somewhat easier because we are leveraging the strength of machine learning, and providing an outcome based on a precise process. However, in cases where the chosen notion of fairness leaves difficult social and moral questions open, such as where to draw the line between public safety and individual fairness, AI should be used more carefully and automation can happen only after clarifying those questions.

Second, as of now, it remains an open question whether AI will have discretionary power. On one hand, the algorithm performs some decision making and choosing one outcome over the other might be viewed as discretionary.[346] On the other hand, the level of intelligence involved in the decision making is minimal, and if only math is behind the algorithm, it could seem like no discretion is involved at all, but in fact the engineers building the system are functionally making many discretionary decisions.[347] The regulation and policy surrounding each one of the policy domains should address this issue. In addition, if by design there is certain room for discretion in the algorithm, the implications of such discretion have to be understood. Some questions that will

[343] *See* Stephen D. Gottfredson & Laura J. Moriarty, *Clinical Versus Actuarial Judgment in Criminal Justice Decisions: Should One Replace the Other?*, 70 FED. PROB. 15, 15–17 (2006) (discussing how probation officers and correctional treatment specialists are affected by universal human decision-making flaws which statistical methods of prediction are not affected by).

[344] Jon Kleinberg et al., *Discrimination in the Age of Algorithms* 10–11 (Nat'l Bureau Econ. Research, Working Paper No. 25548, 2019).

[345] Whittlestone et al., *supra* note 332, at 6.

[346] CHARLES E. HARRIS JR. ET AL., ENGINEERING ETHICS: CONCEPTS AND CASES (5th ed. 2013).

[347] *Id.*

help us determine if AI is discretionary or not include: Given that the set of factors remain the same, and the weight that we give to each factor is also fixed, in regard to person X, will the algorithm always provide the same result? If the answer is yes, then no discretionary power is likely involved. But given that one important component of AI is to learn, the algorithm for the same inquiry today might not lead to the same result a month later. If that is the case, can we say that AI retains discretionary power? How can we know whether the change is the result of learning or of pure inconsistency? And how is this different from inconsistencies in human decision making? Research shows, for example, that fingerprint examiners who were faced with the same evidence at two different times reached different results.[348] In the same manner that we learn to accept inconsistency in human decisions, regardless if they are the result of an error or more educated outcome, will we treat AI decisions similarly, or will we hold AI to a higher standard? Consider another scenario where the factors that the AI considers are known, but the algorithm decides, each separate time, the weight to give to each factor; is this discretion? And will we allow it?

Third, algorithms do not operate in a vacuum, and an interesting question is how the output of the algorithm impacts the administrative discretion of the public agent. If the agent complies with the algorithm's recommendation 100% of the time, there is no effective oversight mechanism. Will we have to decide on a certain percentage of cases in which decision makers are expected to diverge from the algorithm's decision? And how can we know that when the agent diverges from the decision of the algorithm, that it is not in order to reinforce internal bias, but rather out of appropriate exercise of discretion? Regardless of the chosen notion of fairness, the important question is how the final decision of the judge/administrative agent that encapsulate the algorithm's output in it, impacts the individual's right to due process. In order to ensure that due process is respected, transparency as for the way the algorithms was used is required. Both the decision to embrace the outcome or diverge from it could be valid decisions that the judge or administrative agent can take within their discretionary power. However, if the individual does not know in their specific case, if the algorithm was used and how, it will be harder to properly appeal the decision and assess if its fair. Ensuring that the algorithm itself does not reinforce systemic biases and that it leaves ample room for appropriate discretionary action, is an issue that policymakers need to take into account in translating regulation to computer language. Among the advantages that AI offers is increasing responsiveness and enabling the public to hold administrators accountable for their decisions.[349] The accountability mechanisms that are being developed for AI-based algorithms require policymakers to specify the policy explicitly, open it up for public examination,

[348] Itiel E. Dror et al., *Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a "Target" Comparison*, 208 FORENSIC SCI. INT'L 10, 13 (2011).

[349] *See Artificial Intelligence for the American People*, WHITE HOUSE, https://www.whitehouse.gov/ai/ (last visited Jan. 21, 2020) ("The Summit highlighted innovative efforts at Federal agencies that have already adopted AI, and looked ahead to future transformative AI applications that will make government more effective, efficient, and responsive.").

and call for detailing and documenting all steps taken to control discretionary power, so that it might be easier to track and oppose the misuse of discretionary power.[350]

2. Auditing

All notions of fairness discussed in the paper aim to embody fairness in the early stages of the creation of the algorithm; in the design, training, and operation of the algorithm. Auditing takes a bird's-eye view of the process and assesses afterward if the process and outcome are fair. Auditing could be done for direct influence, meaning examining the effect of a specific factor or few factors on the outcome; whether racial considerations affect an arrest decision, for example. Auditing can also be done to detect indirect influence, meaning examining the effect of a specific factor and its proxies.[351] As part of developing a policy for governing the use of algorithms in any policy domain, it is beneficial to write down an auditing checklist for each one of the stages in the life cycle of the algorithm. If done correctly, auditing can help in detecting and proving discrimination. Regulating the process in which algorithms are designed, requiring that all data will be stored and examined, questioning the objectives selected in the training, and examining the outcomes are all auditing methods that can help identify discrimination if it exists, and also help in solving it.[352]

Another advantage of auditing is that it enables policymakers to assess all types of algorithms, including black boxes.[353] Analyzing the output of the algorithm as well as the process, allows policymakers to focus on the important question: whether the algorithm de facto discriminates against a certain minority. Given that governmental institutions usually do not have a practice of collecting high quality data for building and training the algorithm early on, auditing enables continued validation of the algorithm for the particular society, and it deals with any attempts to cheat or overfit the data.[354] Auditing is the only way to prove that the chosen notion of fairness accomplishes its task and that the algorithm is fair toward the individual and the group of individuals it is acting upon.

Legally, auditing can also satisfy the constitutional due process requirement, so long as the process of auditing is detailed, it can help to shed some light on the validity of the algorithm. For example, if we suspect that the algorithm is racially biased or discriminates on the basis of gender, this can be effectively examined in the auditing process.

[350] Thomas J. Barth & Eddy Arnold, *Artificial Intelligence and Administrative Discretion: Implications for Public Administration*, 29 AM. REV. PUB. ADMIN. 332, 336–37 (1999).

[351] Philip Adler et al., *Auditing Black Box Models for Indirect Influence*, 54 J. KNOWLEDGE & INFO. SYS. 95, 96–97 (2017).

[352] Kleinberg et al., *supra* note 344, at 2–3.

[353] James Guszcza et al., *Why We Need to Audit Algorithms*, HARV. BUS. REV. (Nov. 28, 2018), https://hbr.org/2018/11/why-we-need-to-audit-algorithms.

[354] Iyad Rahwan, *Society in the Loop: Programming the Algorithmic Social Contract*, 20 ETHICS & INFO. TECH. 5, 11 (2018).

It is important to mention that auditing as a mechanism cannot stand on its own; it can supplement the chosen notion of fairness and assist in investigating whether this notion of fairness produces good results. Thus, it is recommended that auditing will be implemented in the impact assessment process. Auditing helps in building trust in the algorithm and in starting a conversation about potential changes that need to be made.[355]

D. Encouraging Interdisciplinarity

Perhaps the most important factor to be taken into account both by policymakers and developers is enhancing interdisciplinarity or multi-stakeholder involvement. As this paper has illustrated thus far, the successful creation and implementation of algorithms in our daily life raises countless concerns that could be addressed properly only if experts from different disciplines join forces and bring their unique point of view.[356] Interdisciplinarity in this context means two different things. First, interdisciplinarity means "honoring other expertise" and bringing experts from different disciplines to the table.[357] Second, interdisciplinarity urges all those who were brought to the table to have a basic understanding of the cultural and social background of the problem that the algorithm is intended to solve, and to acknowledge the challenges and limitations of other disciplines that have been more involved in the field longer.[358] For example, in order to build an algorithm that accurately and fairly predicts the risk of recidivism, criminologists, psychologists, prosecutors, defense lawyers, court representatives, correctional officers, computer scientists, and perhaps even other types of professionals need to be involved. It is very important that all participants are familiar with critical and non-trivial perspectives such as black, feminist, and queer views.[359]

VII. CONCLUSION

The statistician George Box is known for his famous quotation: "All models are wrong, but some are useful."[360] This quote is used in statistics classes to teach students that every model is, in some sense wrong because it represents a simplified version of reality. Each model has its own downside, it ignores certain aspects of reality, or it aims to equalize one side of the equation.[361] The goal of this paper is to offer a translation of this quote into something like, "most

[355] Jessi Hempel, *Want to Prove Your Business is Fair? Audit Your Algorithm*, WIRED (May 9, 2018), https://www.wired.com/story/want-to-prove-your-business-is-fair-audit-your-algorithm/.

[356] MEREDITH WHITTAKER ET AL., AI NOW REPORT 2018 36–37 (Dec. 2018), https://ainowinstitute.org/AI_Now_2018_Report.pdf.

[357] *Id.* at 36.

[358] *Id.*

[359] Paul Dourish et al., *Reflective HCI: Towards a Critical Technical Practice*, Proceedings of CHI '04 Extended Abstracts on Human Factors in Computing Systems 1727, 1727–28 (Apr. 2004) https://doi.org/10.1145/985921.986203.

[360] Edward H. Field, "*All Models Are Wrong, But Some Are Useful,*" 86 SEISMOLOGICAL RES. LETTERS 291, 291 (2015).

[361] *Id.* at 292.

models are right, but it depends how we use them."[362] There are many notions of fairness and each one corresponds with a different legal mechanism which makes it suitable for solving a certain pressing social problem. But a solution that is applicable for credit scoring might not be applicable for criminal justice. Similarly, an algorithm that leads to fairer results in college admission cannot necessarily be used to solve problems in hiring. The three main categories of fairness notions aim to protect different objects and their compatibility with different legal mechanisms vary.[363] The gap between the different disciplines that are studying fairness and the conceptual differences between them can lead to some confusion. For example, certain notions that might be dismissed quickly by computer scientists due to some tradeoffs in the accuracy could be suitable for some legal and policy domains where we acknowledge that efficiency and accuracy alone are not the end of the game. In other cases, solutions that might seem technically perfect could be hard to implement from a legal perspective due to the need to determine very sensitive questions that are difficult to reach consensus on. There is a conceptual difference between the way law and computer science operates, which requires the two disciplines to make some compromises in order to benefit society overall. Therefore, we should be very cautious before making any general conclusion about one fairness notion or another.

The typology presented in this paper is meant to encourage research to focus on the context. Such approaches have been proven useful when technology and law are at an intersection.[364] AI algorithms cannot replace social or legal reforms that need to be made in order to cultivate a more just society, but collaboration between all actors in the field can at least ensure that we are on the right path.

[362] *See supra* Section VI.B.1 (discussing how the choice of fairness approach for each situation is critical).

[363] *See supra* Section I (describing the three main fairness categories).

[364] Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 154 (2004).