NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Ekaterina Semenova, Ekaterina Perevoshchikova,

Alexey Ivanov, Mikhail Erofeev

FAIRNESS MEETS MACHINE LEARNING: SEARCHING FOR A BETTER BALANCE

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LAW

WP BRP 93/LAW/2019

Ekaterina Semenova,[1] Ekaterina Perevoshchikova,[2]

Alexey Ivanov,[3] Mikhail Erofeev[4]

FAIRNESS MEETS MACHINE LEARNING: SEARCHING FOR A BETTER BALANCE

Machine learning (ML) affects nearly every aspect of our lives, including the weightiest ones such as criminal justice. As it becomes more widespread, however, it raises the question of how we can integrate fairness into ML algorithms to ensure that all citizens receive equal treatment and to avoid imperiling society's democratic values.

In this paper we study various formal definitions of fairness that can be embedded into ML algorithms and show that the root cause of most debates about AI fairness is society's lack of a consistent understanding of fairness generally. We conclude that AI regulations stipulating an abstract fairness principle are ineffective societally.

Capitalizing on extensive related work in computer science and the humanities, we present an approach that can help ML developers choose a formal definition of fairness suitable for a particular country and application domain.

Abstract rules from the human world fail in the ML world and ML developers will never be free from criticism if the status quo remains. We argue that the law should shift from an abstract definition of fairness to a formal legal definition. Legislators and society as a whole should tackle the challenge of defining fairness, but since no definition perfectly matches the human sense of fairness, legislators must publicly acknowledge the drawbacks of the chosen definition and assert that the benefits outweigh them. Doing so creates transparent standards of fairness to ensure that technology serves the values and best interests of society.

JEL Classification: K19

Keywords: Artificial Intelligence; Bias; Fairness; Machine Learning; Regulation; Values; Antidiscrimination Law;

[1] Research Fellow at HSE – Skolkovo Institute for Law and Development at the National Research University – Higher School of Economics, Moscow. Email: evsemenova@hse.ru

[2] Research Fellow at HSE – Skolkovo Institute for Law and Development at the National Research University – Higher School of Economics, Moscow. Junior Research Fellow at University College London, UK. PhD candidate at the University of Southampton, UK. Email: eperevoshchikova@soton.ac.uk.

[3] Director at HSE – Skolkovo Institute for Law and Development at the National Research University – Higher School of Economics, Moscow. LLM, Harvard University. Email: aivanov@hse.ru

[4] Research Fellow at the Intellectual Information Technologies Lab at the Department of Computational Mathematics and Cybernetics at Lomonosov Moscow State University, Moscow. PhD in Computer Science. Email: merofeev@graphics.cs.msu.ru

1. Introduction

To decide whether an accused individual should be eligible for bail or should be held in custody pending trial, judges review multiple factors. Currently, US judges assess, among other things, reports from COMPAS—an algorithm-based machine-learning (ML) system for scoring criminal defendants according to their risk of reoffending. In 2016, a fascinating criminal-justice controversy broke out when a Wisconsin circuit court charged Eric Loomis in relation to a drive-by shooting. Dissatisfied with the results of his risk score, Loomis appealed to the Wisconsin Supreme Court.[5]

Meanwhile, COMPAS came under the scrutiny of the investigative media organization ProPublica, which found that the system classified black defendants as medium or high risk much more often than white defendants. The organization determined that the result was harsher treatment of black defendants who ultimately avoided recidivism.[6] Northpointe, which developed the ML system, responded that the algorithm was fair because the ratio of those who received high risk scores and then reoffended was essentially the same (around 60%) for both blacks and whites.[7] Therefore, it argued, the scores mean the same thing regardless of race. The Wisconsin Supreme Court ultimately ruled in favor of using COMPAS.

This controversy reveals an interesting point: there is no single approach to fairness in ML. ProPublica argued that the COMPAS algorithm was unfair to black defendants, and its findings were compelling. But the same is true of Northpointe, which argued that COMPAS is indeed fair. These two understandings of fairness are not reconcilable.

Choosing the correct approach to ML fairness is an upstream topic for computer science and the humanities. Whereas computer scientists are developing methods to embed fairness into ML algorithms, legal researchers are attempting to find a more or less universal understanding of fairness and translate formal computer-science-based definitions into law. As Berk et al. (2018) note, the multidisciplinary conversation about translating fairness from machine language to public policy, and vice versa, is in its early stages, and it needs greater attention.[8]

Fairness is a pivotal principle underpinning the democratic and social values of our society. In his theory of justice as fairness, John Rawls posits that society must assure each

[5] State of Wisconsin v Eric Loomis, 881 N.W.2d 749 (Wis. 2016).
[6] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, 'Machine Bias' (2016) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[7] 'A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear' (2016) https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?noredirect=on
[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth, 'Fairness in Criminal Justice Risk Assessments: The State of the Art' Sociological Methods & Research (2018) https://doi.org/10.1177/0049124118782533

citizen "an equal claim to a fully adequate scheme of equal basic rights and liberties."[9] Democratic constitutions guarantee citizens equal rights and opportunities as a necessary premise for the functioning of society.

Technological advances, including the rise of artificial intelligence (AI), raise this old question anew. Who should be responsible for making algorithms fair? Should tech giants be solely responsible? Or should regulators and courts have a say? Crucially, how can we reconcile abstract legal rules on fairness, such as the idea that every algorithm should be fair, with the logic of these algorithms?

This paper seeks to answer the question of whether stipulating fairness as an abstract guiding principle is enough to provide a high standard of equality and to preserve democratic values in a world powered by algorithms. To answer this question, we focus primarily on the North American antidiscrimination paradigm and US case law, since it provides a rich trove of cases and arguments on fairness and discrimination,[10] and with regard to emerging technologies such as AI. Without making any assumptions on the supremacy of the US approach to fairness compared to European or Russian ones, we find it illustrative of the strengths and shortcomings of the current fairness notion. We believe that we can draw a useful example from this study of US case law that will be helpful in European and Russian contexts.

Capitalizing on extensive related work in both computer science and the humanities,[11] we suggest a novel approach to dealing with the mismatch between abstract social definitions of fairness and precise ML definitions. We argue that an alternative to an *abstract* legal definition is a *formal* legal definition. Legislators and society as a whole should confront the challenge of defining fairness, but since no definition perfectly matches the human sense of fairness, legislators must publicly acknowledge the drawbacks of the chosen definition and assert that the benefits outweigh them. Doing so increases legal transparency and creates quality standards of fairness.

2. The Call for a Fair Machine-Learning Algorithm

[9] Para 4.3. in 'John Rawls' in the Stanford Encyclopedia of Philosophy https://plato.stanford.edu/entries/rawls/#JusFaiJusWitLibSoc; 'John Rawls' in the Internet Encyclopedia of Philosophy https://www.iep.utm.edu/rawls/#SH2d.

[10] For the overview of antidiscrimination law cases see, for example, Philip F. Rubio 'A History of Affirmative Action 1619 – 2000' University Press of Mississippi; Abigail Nurse 'Anti-Subordination in the Equal Protection Clause: A Case Study' 89 N.Y.U. L. Rev. 293 (2014).

[11] To dive into relevant work on AI fairness, please, refer to Solon Barocas and Andrew D. Selbst, 'Big data's disparate impact.' Calif. L. Rev. 104 (2016): 671; Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, 'Accountable Algorithms' 165 U. Pa. L. Rev. 633 (2017) https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

The hype about and the misunderstanding of ML's power and workflow have resulted in new regulations and ethical guidelines that aim to make software developers build fair, accountable, and transparent algorithms. Fairness is a cornerstone of ML algorithms that aim to ensure accountability and transparency.[12]

Both general human-rights regulations—such as Article 21 of the Charter of Fundamental Rights of the European Union, Article 14 of the European Convention on Human Rights and Article 1 of the American Convention on Human Rights—and ML regulations place a fairness requirement at the forefront of decision-making. For example, the European Union's General Data Protection Regulation (GDPR) demands that algorithms embed fairness in the form of equal and non-discriminatory treatment of people.[13] In Recital 71, the GDPR demands that automated decision-making and profiling algorithms "implement technical and organizational measures" to "prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect." It also requires that persons should have the right to refuse subjection to automated decision-making without human intervention. This opt-out regime—meaning a person may switch from automatic mode to manual mode—reduces the potential impact of unfair decisions by an AI algorithm, but it fails to eliminate the problem of AI fairness, because making all decisions in manual mode is unfeasible.

Another draft regulation on algorithm fairness is the US Algorithmic Accountability Act of 2019.[14] This bill requires companies to audit ML algorithms and assess whether they are fair. Companies must conduct studies on an algorithm's impact to evaluate, among other things, fairness and the risks that unfair decisions pose to consumers. Violators would be subject to penalties.

The European Commission's recent Ethics Guidelines for Trustworthy AI addresses fairness as one of several ethical imperatives and assigns responsibility for complying with fairness to developers.[15]

---

[12] European Government in its study 'A Governance Framework for Algorithmic Accountability and Transparency' considers fairness as "a fundamental component underpinning responsible systems and suggest that algorithmic processes should seek to minimise their potential to be unfair and maximise their potential to be fair. Transparency and accountability provide two important ways in which this can be achieved."
http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf

[13] Bryce Goodman, Seth Flaxman, 'European Union regulations on algorithmic decision-making and a "right to explanation"
https://arxiv.org/pdf/1606.08813.pdf
Bryce Goodman, 'A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection' http://www.mlandthelaw.org/papers/goodman1.pdf

[14] 'Algorithmic Accountability Act of 2019'
https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf

[15] European Commission, 'The Ethics Guidelines for Trustworthy Artificial Intelligence' (April 2019) 13
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Diversity

The Declaration on Ethics and Data Protection in Artificial Intelligence, issued by the International Conference of Data Protection and Privacy Commissioners in 2018,[16] stresses the significance of "elaborating specific guidance and principles in addressing biases and discrimination". The IEEE, a world leader in the creation of technical standards addresses the matter in its recent "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems." This document encourages the development of national and global policies regarding the ethical principles and standards for ML algorithms; it argues that embedding social values in ML algorithms is crucial.[17] The IEEE, however, says that society has yet to provide universal guidelines for this task.[18] It is correct with regard to fairness: no universal notion of fairness exists, and both the humanities and STEM fields have produced contradictory definitions.[19]

Alongside government regulations on AI ethics, companies have adopted their own principles.[20] These codes of conduct for fair AI manifest a company's adherence to equitable and accountable use of the technology.

Ethical regulations and codes of conduct, however, leave open the issue how to implement a policy in AI. Legislators have demanded that ML algorithms encode fairness, and companies hasten to declare their compliance. These legislators, however, identify no precise features that developers can build into algorithms, keeping the fairness requirement abstract.

To create fair ML algorithms, developers must understand what fairness is when choosing an appropriate definition to embed in these algorithms. That is, they should have clear criteria for assessing whether an algorithm is fair. Therefore, society needs universal fairness guidelines so that ML developers can follow a unified approach when designing algorithms rather than simply choosing one among many. Otherwise, an ML algorithms will be fair according to one definition while being unfair according to all others.

---

[16] 40th International Conference of Data Protection and Privacy Commissioners, 'Declaration on Ethics and Data Protection In Artificial Intelligence' (October 23, 2018) https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf

[17] IEEE 'Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems'. 33-54 https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

[18] ibid. 33.

[19] For different approaches to fairness understanding in humanities see, for example, Jack M. Balkin, Reva B. Siegel, 'The American Civil Rights Tradition: Anticlassification or Antisubordination?' 58 U. Miami L. Rev. 9 (2003) https://repository.law.miami.edu/cgi/viewcontent.cgi?article=1425&context=umlr
For formal notions of fairness developed in STEM refer to Sahil Verma, Julia Rubin, 'Fairness Definitions Explained' http://fairware.cs.umass.edu/papers/Verma.pdf
Subsequent sections of this paper focus on different notions of fairness developed in both humanities and STEM.

[20] Sundar Pichai, 'AI at Google: our principles' https://www.blog.google/technology/ai/ai-principles/
Microsoft FATE: Fairness, Accountability, Transparency, and Ethics in AI https://www.microsoft.com/en-us/research/group/fate/ https://www.ibm.com/blogs/watson/2018/07/trust-in-the-age-of-ai-build-fairness-into-machine-learning-models/

Because it is a social value, formal notions of fairness embedded in ML algorithms are grounded in the humanities. Below, we focus on fairness both in the law and in ML to show how inconsistent understandings in the humanities lead to disputes over fairness in ML.

3. Fairness as a Legal Principle

Fairness is a fundamental legal principle.[21] For centuries, fairness—together with other legal principles such as transparency, accountability, and proportionality—has been an abstract rule open to interpretation and judicial discretion.

The notion of fairness has changed over time. Furthermore, as Google's principles on AI note, "distinguishing fair from unfair biases is not always simple and differs across cultures and societies."[22] *Ius est ars boni et aequi,* said ancient Roman law, despite treating slaves as property. English law punished homosexuality, and only the Equality Act 2010 fully protected lesbian, gay, bisexual, and transgender rights. Some countries, however, still punish or discriminate against individuals on the basis of sexual orientation. Other examples showing that fairness depends on social context abound. The abstract nature of rules for fairness, however, stagnated over the centuries, leaving room for interpretation according to societal needs. "Judges look not to rules, but to fairness first, and then, once they have perceived the just solution, they rationalize it via case opinion."[23]

That said, fairness has been always defined through the abstract principles of how society should function. There are generally two philosophical approaches to fairness: equality of outcome and equality of opportunities. Equality of opportunities means that "agent(s) […] have a chance to attain a specified goal without the hindrance of some obstacle".[24] Equality of outcome instead means that all members of population share the same outcome (e.g. wealth or education), not just have a chance to achieve this outcome.[25] The philosopher John Rawls epitomized the substantive equality of opportunity approach. In his theory of "justice as fairness" he identifies two principles of fairness. The first principle holds that society must assure each citizen "an equal claim to a fully adequate scheme of equal basic rights and liberties."[26] The second principle refers to the distribution of goods, wealth and opportunities. According to this, social structures must satisfy the requirements of "fair

[21] Jeremy Miller, 'The Science of Law: The Maturing of Jurisprudence into Fundamental Principles in Fairness' Western State University Law Review, Vol. 13, No. 2, 367 (1986) https://ssrn.com/abstract=925947
[22] Sundar Pichai (n 20) Op.cit.
[23] Jeremy Miller (n 21) Op.cit.
[24] 'The Concept of Equality of Opportunity' https://edeq.stanford.edu/sections/concept-equality-opportunity
[25] See 'Equality of Outcome', https://edeq.stanford.edu/sections/equality-outcome.
[26] Para 4.3 in 'John Rawls' in the Stanford Encyclopedia of Philosophy https://plato.stanford.edu/entries/rawls/#JusFaiJusWitLibSoc; 'John Rawls' in the Internet Encyclopedia of Philosophy https://www.iep.utm.edu/rawls/#SH2d.

equality of opportunity." However, Rawls admits that differentiation is inevitable and holds that "social and economic inequalities […] are to be to the greatest benefit of the least advantaged members of society",[27] thereby bringing substantive criteria into the understanding of fairness. Equality of opportunity is a pre-dominant principle of modern society, though how exactly it should be implemented is open to interpretation.[28]

Fairness is a social value, and society should decide what it means, but neither philosophers nor legal scholars have produced a single approach to fairness, so developers are left to choose between different options when attempting to develop fair algorithms. Fairness as a legal principle is kept deliberately broad so judges and administrators can use their discretion in addressing the variety of real-life cases yet still follow specific practices and procedures. The question is whether AI will challenge this status quo and bring a more concise, positivistic view of fairness to the law.

To answer this question, we focus on how the legal community and judges understand fairness. Below, we provide a brief overview of case law revealing the approaches to fairness developed by legal scholars. This overview shows that despite more than 50 years of antidiscrimination law, scholars have failed to develop a notion of fairness that courts can follow.

The root cause of AI fairness issues is the different understandings of fairness in society. Moreover, courts even in the same country are inconsistent in their choice of fairness doctrine. We illustrate inconsistencies in the understanding of fairness by studying US case law, as problems of fairness and equality have garnered extensive attention from scholars and society at large owing to that nation's history of social stratification.

Though the concept of fairness has been extensively explored by many ancient and modern philosophers and has been closely intertwined with the notion of justice, the US has become a major jurisdiction where the fairness principle has been applied scrupulously in employment, education, criminal sentencing, and other domains.[29] Thus, US case law is more mature in this regard than that of other regions. For example, the European Court of Human Rights decided its first case on school desegregation in 2007, 60 years after the US Supreme

[27] John Rawls, *Justice as Fairness (2001) 42-43, quoted in* 'John Rawls' in the Stanford Encyclopedia of Philosophy.
[28] For example, some scholars state that taking equality of opportunity seriously with respect to race, gender and political representation means that we should care about and expect equality of outcome (see Anne Phillips, 'Defending Equality of Outcome' Journal of Political Philosophy, Journal of political philosophy, 12, no. 1 (2004): 1-19).
[29] Grainne de Burca, 'The Trajectories of European and American Antidiscrimination Law' American Journal of Comparative Law Vol. 60, No. 1 (2012) https://ssrn.com/abstract=1950697
Aaron Cohen, 'Culture, Values, and Organizational Fairness' In: Fairness in the Workplace. Palgrave Macmillan, London (2015); Richard B. Darlington, 'Another Look at "Cultural Fairness"' Journal of Educational Measurement 8, no. 2 (1971): 71-82; Tae-Yeol Kim, Kwok Leung, 'Forming and Reacting to Overall Fairness: A Cross-Cultural Comparison' Organizational Behavior and Human Decision Processes (2007) https://doi.org/10.1016/j.obhdp.2007.01.004

Court's pioneering case.[30] We therefore discuss mainly the case law and fairness paradigm of the US.

Researchers describe two high-level understandings of fairness in the US: anticlassification and antisubordination.[31] According to Balkin and Siegel, anticlassification "holds that the government may not classify people either overtly or surreptitiously on the basis of a forbidden category: for example, their race."[32] Antisubordination holds that "it is inappropriate for certain groups in society to have subordinated status because of their lack of power in society as a whole. This approach seeks to eliminate the power disparities between men and women, and between whites and non-whites, through the development of laws and policies that directly redress those disparities."[33]

The application of these two approaches—anticlassification and antisubordination—"shifts over time in response to social contestation and social struggle,"[34] and as US case law shows, neither the Supreme Court nor the lower courts are consistent in their understanding of fairness.[35] To illustrate this, we discuss notable US court decisions related to fairness and equality, and we identify the doctrine on which they are based.

The landmark US Supreme Court case *Griggs v. Duke Power Co.* (1971) is an example of the moderate antisubordination approach.[36] In this case Duke Power introduced a policy under which it only considered employees having a high school diploma for promotion, thereby limiting the eligibility of black employees.

Under the anticlassification approach, Duke Power's policy should be legal and fair, since promotion decisions were not based on race or any other protected class, but on education. The Supreme Court, however, followed a moderate antisubordination approach, deciding that the promotion policy is illegal because of its disparate impact and because it lacks a business-related motivation.[37] In deciding this case, the court conditionally prohibited disparate impact, whereas antisubordination requires unconditional denial of disparate-impact practices.

---

[30] Kristine Bowman and Jiri Nantl, 'Liability and Remedies for School Segregation in the United States and the European Union' International Journal of Education Law and Policy (2015) https://ssrn.com/abstract=2675199

[31] See, for example, Jack M. Balkin, Reva B. Siegel, 'The American Civil Rights Tradition: Anticlassification or Antisubordination?' 58 U. Miami L. Rev. 9 (2003) 11 https://repository.law.miami.edu/cgi/viewcontent.cgi?article=1425&context=umlr

[32] ibid. P. 10.

[33] Ruth Colker, 'Anti-Subordination Above All: Sex, Race, And Equal Protection' NYUL Rev. 61 (1986) 1006

[34] Jack M. Balkin, Reva B. Siegel (n 31) 10

[35] Abigail Nurse claims that both The Supreme Court and lower courts have failed to consistently embrace either anticlassification or anti-subordination to interpret the Equal Protection Clause. See Abigail Nurse, 'Anti-Subordination In The Equal Protection Clause: A Case Study' (2014) N.Y.U. L. Rev.

[36] Kimberly A Yuracko, 'Gender Nonconformity and the Law' New Haven; London: Yale University Press (2016) 55.

[37] Sam Corbett-Davies and Sharad Goel, 'The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning' (2018) 4-5 https://arxiv.org/pdf/1808.00023.pdf

Two other landmark cases, *Washington v. Davis* (1976)[38] and *The City of Los Angeles Department of Water and Power v. Manhart* (1978),[39] illustrate anticlassification. In *Washington v. Davis,* two black claimants challenged The District of Columbia Police Department's recruiting policy, alleging that its mandatory written test discriminated against black applicants. The court, however, ruled that the hiring policy is legal because it is not based on the applicant's race and because the test is a neutral proficiency tool. Thus, it applied the anticlassification approach by deciding that the policy does not reference protected classes and thus avoids discrimination.

The City of Los Angeles Department of Water and Power v. Manhart is another example in which the court followed anticlassification—this time, regarding retirement-plan payments that were based on an employee's gender. In this case The City of Los Angeles Department of Water forced female employees to make larger payments to the employee retirement plan than male employees because they have a longer life expectancy. The court found the retirement plan illegal because it discriminates against a protected class (gender).

Fisher v. The University of Texas (2016)[40] illustrates the antisubordination approach: the court ruled that the university has the right to enforce racial diversity through an approved university enrollment policy that takes into account the applicant's race, arguing that this policy does not lead to discrimination. The court's ruling said that "race-conscious affirmative action programs for college admissions are legally permissible to further the government's interest in promoting diversity."[41]

In *The State of Wisconsin v. Eric Loomis* (2016),[42] the court stuck with the anticlassification approach, deciding that the use of a risk-assessment algorithm is legal and does not violate an individual's right to due process.

To summarize, US courts inconsistently decide cases that involve fairness, choosing between the anticlassification and antisubordination approaches. Society, therefore, cannot adhere to just one fairness doctrine because the sensitive nature of the topic leads to conflicting opinions. As we show in the next section, this inconsistency prevents the implementation of a strict, formal fairness definition in ML algorithms; furthermore, it makes impossible the development of ML algorithms that all parties consider fair.

4. Embedding Fairness into an Algorithm

[38] Washington v. Davis, 426 U.S. 229 (1976)
[39] City of Los Angeles Department of Water & Power v. Manhart, 435 U.S. 702 (1978)
[40] Fisher v. University of Texas at Austin, 579 U.S. (2016)
[41] Sam Corbett-Davies, Sharad Goel, Op.cit.
[42] State of Wisconsin v Eric Loomis, 881 N.W.2d 749 (Wis. 2016).

For ML algorithms, fairness is not the default.[43] Moreover, requiring that an algorithm "be fair" is insufficient, since algorithms need formal definitions of the values on which they are built.

ML is a science that seeks to optimize different quality metrics. The algorithms can be trained to satisfy only objectives with a formal (mathematical) definition, not abstract definitions. Therefore, the human understanding of fairness requires expression in formal terms before it can become a goal of AI.

So far, computer scientists have produced at least 21 definitions of fairness.[44] To illustrate the connection between debates about fairness in general and debates about fairness in AI, we briefly review the five formal definitions on the continuum between anticlassification and antisubordination. For simplicity, we express each one in terms of an imaginary classifier that, given certain information about a criminal, outputs the score expressing probability of recidivism.

1. **Fairness through unawareness.** Regulations must not discriminate on the basis of gender, race, age, religion, health, or any other individual characteristic (protected classes). A seemingly straightforward way to mitigate the ML-algorithm bias is to avoid using data on protected classes when training the algorithm. In other words, the algorithm should be unaware of sensitive information about individuals. Applied to the problem of recidivism, this fairness-through-unawareness paradigm means the classifier will have no access to information about a criminal's race, gender, age, and so on. But many cases have shown that an algorithm can make inferences about protected classes on the basis of other nonprotected attributes.[45] Thus, fairness through unawareness fails.
2. **Calibration.** An ML algorithm is said to be calibrated if the probability of outcome A is independent of protected class, given that the algorithm's estimated probability for outcome S is equal to s. Thus, the following equation holds: $P(A|S=s, B=b_1) = P(A|S=s, B=b_2)$, where $B=b_1$ and $B=b_2$ denote attribution to a certain protected class, and $P(A|B)$ is the conditional probability of outcome A given that event B has occurred. For example, among those classified as having a high risk of recidivism, the portion who will actually commit another crime

[43] Josh Lovejoy, 'Fair Is Not the Default,' (February 2018) https://design.google/library/fair-not-default/
[44] Songül Tolan, 'Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges' JRC Digital Economy Working Paper (October 2018) 8.
[45] Alexey Romanov et al., 'What's in a Name? Reducing Bias in Bios without Access to Protected Attributes.' arXiv preprint arXiv:1904.05233 (2019) https://arxiv.org/abs/1904.05233
Solon Barocas and Andrew Selbst, 'Big data's disparate impact.' Calif. L. Rev. 104 (2016): 671.

should be the same for both white and black people. As Barocas et al. note, ML algorithms usually maintain this type of fairness by default.[46]

3. **Equality of opportunity.** Initially proposed by Hardt et al.,[47] the equality-of-opportunity paradigm imposes constraints on "advantaged" outcomes (in most decision-making problems, one of two possible outcomes is clearly advantageous—e.g., not defaulting on a loan, satisfying a college's admission requirements, and not committing a crime). In particular, it requires that the probability of a *correct* classification for the individual with an "advantaged" outcome be independent of protected class. The following equation should therefore hold: $P(\hat{A} \ | A, B = b_1) = P(\hat{A} | A, B = b_2)$, where A means that the "advantaged" outcome actually occurs and $\hat{A}$ means the classifier predicted that the "advantaged" outcome will occur. $B = b_1$ and $B = b_2$ denote attribution to a certain protected class. For example, given an individual who *will not* commit further crimes, the probability of a pardon should be independent of protected class.
4. **Equalized odds.** Equalized odds is stricter a form of the equality-of-opportunity paradigm. Besides constraints on "advantaged" outcomes, it also imposes constraints on "disadvantaged" outcomes. In particular, it additionally requires that the probability of *correctly* classifying an individual with a "disadvantaged" outcome be independent of protected class. For example, given an individual who *will not* commit further crimes, the probability of a pardon should be independent of protected class *and* the probability of jail time for an individual who *will* commit further crimes should be independent of protected class.
5. **Demographic parity.** The demographic-parity paradigm requires that the probability of the classifier assigning outcome A be independent of protected class: $P(\hat{A} \ | B = b_1) = P(\hat{A} \ | B = b_2)$, where $\hat{A}$ means the classifier predicted that outcome A will occur. For example, this paradigm requires that the classifier assign the "low risk of reoffending" label to the same fraction of white defendants as black defendants.

All these paradigms in some sense align with the human understanding of fairness. Unfortunately, however, we cannot combine all of them into one ML method. For example,

[46] Solon Barocas, Moritz Hardt, Arvind Narayanan, ,'Fairness and Machine Learning' (2018) http://www.fairmlbook.org
[47] Moritz Hardt, Eric Price, and Nati Srebro, 'Equality of opportunity in supervised learning.' Advances in neural information processing systems (2016).

Chouldechova[48] shows that the classifier cannot both undergo calibration and satisfy the equalized-odds paradigm if the base rate (e.g., the recidivism rate) differs among protected groups. Access to protected-class data is necessary to satisfy the equality-of-opportunity, equalized-odds, and demographic-parity constraints.[49] Such access clearly contradicts the fairness-through-unawareness principle. Moreover, few scenarios are able to combine calibration and equality of opportunity. For example, as Pleiss et al. show, improving the COMPAS classifier to satisfy both the calibration and equality-of-opportunity paradigms is impossible.[50]

Owing to the incompatibility of fairness paradigms, any ML system that makes predictions about individuals is, in some sense, unfair because it necessarily fails to satisfy at least one fairness paradigm. Therefore, one can always argue that a particular ML algorithm is unfair despite the developer's efforts. We claim, however, that creating a fair ML algorithm is not a computer-science issue, since the algorithm can satisfy any feasible formal definition of fairness. The root of this problem is the absence of a consistent societal understanding of fairness. To support this claim, first consider the connection between the fairness paradigms we describe in this section and the anticlassification and antisubordination understandings we described in the previous section:

1. Fairness through unawareness can be considered a radical anticlassification approach to fairness. The classifier cannot base its decision on protected classes because it has no access to them. Moreover, any disparate impact on protected groups due to the classifier's use of proxy attributes (e.g., neighborhood) that are well correlated with protected attributes does not contradict anticlassification, since as Balkin and Siegel[51] note, disparate impact is legal under that approach.
2. Calibration can be considered an anticlassification approach to fairness, because as mentioned, most classifiers are calibrated by default and calibration seldom requires access to protected-class information during the deployment stage. But verifying the calibration requires access to protected classes during development. Therefore, this paradigm can be regarded as on or near the boundary between anticlassification and antisubordination.

[48] Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments' (2016) https://arxiv.org/pdf/1610.07524.pdf
[49] Kenneth Holsteinet al., 'Improving fairness in machine learning systems: What do industry practitioners need?' arXiv preprint arXiv:1812.05239 (2018) https://arxiv.org/abs/1812.05239; Indrė Žliobaitė, Bart Custers ,'Using sensitive personal data may be necessary for avoiding discrimination in data- driven decision models' Artificial Intelligence and Law 24(2): 183–201 (2016)
[50] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, Kilian Q. Weinberger, 'On Fairness and Calibration' NIPS (2017) http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf
[51] Jack M. Balkin, Reva B. Siegel (n 31) Op. cit.

3. Equality of opportunity and equalized odds are shades of antisubordination. As mentioned, compliance with these paradigms often requires access to protected classes during deployment. As Hardt et al.[52] note, their implementation can be regarded as affirmative action. Notably, affirmative action is legal only under the antisubordination understanding of fairness.[53]
4. Demographic parity is an example of radical antisubordination, since it eliminates stratification based on protected class while disregarding possible base-rate differences among protected groups. For example, despite different recidivism rates among protected groups, this approach classifies the same fraction of each group as potential reoffenders.

Because it is impossible to simultaneously satisfy different formal fairness paradigms and conform to different understandings of fairness, the absence of a consistent societal understanding (e.g., the Supreme Court's inconsistent use of antisubordination versus anticlassification) naturally leads to the impossibility of developing an ML algorithm that all members of society will consider fair.

To further support our claim that fairness in ML is not an ML-specific issue for computer scientists to resolve, consider the following thought experiment. A judge decides whether the accused is likely to break the law again. One can check that judge's record as to whether the decisions satisfy the calibration and equalized-odds paradigms. Since these two paradigms are incompatible, however, the judge has no chance[54] of being "fair" under both.

Because the ML-fairness issue has deep roots, laws that require ML algorithms to comply with some abstract fairness principle are unworkable, since they fail to define any particular objective that ML developers should strive for. Moreover, the absence of a formal legal definition of fairness, despite several potentially acceptable possibilities, exacerbates the issue of ML fairness, as it allows double standards and makes the legal framework nontransparent. In particular, a party can always adopt the most beneficial definition and claim that a given ML algorithm is fair (or unfair) because it satisfies (or fails to satisfy) that definition. Thus, AI developers have no way to protect themselves from accusations that their algorithms are unfair.

An illustration of this uncertainty is COMPAS, described in the introduction. The extensive media attention it garnered[55] was based on ProPublica's finding that COMPAS is

[52] Moritz Hardt, , Eric Price, and Nati Srebro (n 47) Op.cit.
[53] Jack M. Balkin, Reva B. Siegel (n 31) Op. cit.
[54] Except for the impossible scenario in which the judge never makes mistakes when predicting the future.
[55] 'A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear' 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=.cb9347d04be7

unfair because it fails to comply with ProPublica's definition of fairness—COMPAS yields different false-positive rates for different races (i.e., it violates the equality-of-opportunity paradigm).[56] In response to this finding, the developer claimed the algorithm is indeed fair because it is calibrated, that is, among those whom COMPAS predicts will be reoffenders, the actual recidivism rate is the same across races.[57] In the absence of a formal legal definition of fairness, determining whether COMPAS complies with the law is impossible, since it is fair by one definition but unfair by another.

The only way to eliminate this uncertainty is to establish a single formal definition of fairness in the law, or at least define fairness for each domain (e.g., calibration for sentencing and equality of opportunity for hiring). Such an outcome, however, is unlikely in the near term owing to the intractability of the moral and ethical questions surrounding this topic.

We think the only way for ML developers to reduce this uncertainty is to choose a fairness paradigm by analyzing the dominant approach (anticlassification versus antisubordination) of a particular domain and country. Notably, no single size fits all, since an ML algorithm that is fair and legal in one country could be illegal in another. For example, a university-admissions classifier that satisfies the equality-of-opportunity paradigm may be legal in the US but illegal in the UK, since UK law prohibits affirmative action.[58]

To conclude this section, we note that the choice of fairness paradigm is not an issue that ML developers must address when creating a fair algorithm. Numerous other fairness-related issues arise, for example, how to collect a diverse and unbiased training data set. But these questions are beyond the scope of our paper and we instead focus exclusively on how to formally define fairness, since it is the most general question and, as we have shown, it is not a computer-science issue, but a legal and social one. In the conclusion we offer recommendations for legislators and ML developers to help tackle this problem.

5. Conclusion

Properly defining fairness is a tough issue for both the humanities and computer science—mostly because bias is rooted in society. A vivid example is word embedding: for example, society tends to associate *software engineer* with men rather than women, but it tends to

[56] Ed Yong 'A Popular Algorithm Is No Better at Predicting Crimes Than Random People' (2018) https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/?single_page=true

[57] Songül Tolan, 'Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges' JRC Digital Economy Working Paper 2018-10.

[58] For further analysis of the affirmative action illegality in the UK, please refer, for example, to Uduak Archibong and Phyllis W. Sharps, 'A Comparative Analysis of Affirmative Action in the United Kingdom and the United States' Journal of Psychological Issues in Organizational Culture (2013) https://onlinelibrary.wiley.com/doi/pdf/10.1002/jpoc.21094 Steven Teles, 'Why is there no affirmative action in Britain?' American Behavioral Scientist 41, no. 7 (April 1998) 1004–26. doi:10.1177/0002764298041007010.

associate *homemaker* with women rather than men.[59] To ensure fairness, "one should attempt to debias society rather than word embeddings."[60] Nevertheless, society must find ways to ensure fairness.

Throughout history, society has always lacked a unified approach to fairness. Neither philosophers nor legal scholars and judges have developed such an approach; they waiver between contradictory definitions of fairness. To date, society has followed a case-by-case approach to fairness via case law, when judges intervene and restore justice. The abstract nature of fairness rules, which have gone unchanged for centuries, gave judges flexibility to interpret fairness according to societal needs. This *ex post* approach to restoring fairness via case law, however, is less effective in ML owing to the technology's near ubiquity and its large-scale application in numerous domains.

Furthermore, recent documents on AI prioritize the role of companies and developers in creating fair ML algorithms, demanding that they decide from the outset what fairness is and which definition to employ. For example, the US Algorithmic Accountability Act of 2019 makes developers responsible for ML algorithms being fair. The European Commission's recent Ethics Guidelines for Trustworthy AI requires that AI operators develop, deploy, and use ML algorithms in a manner that adheres to the ethical principle of fairness, but it acknowledges that finding the right solution based on the fairness principle is difficult.[61] To ensure fairness, the guidelines recommend that developers "approach ethical dilemmas and trade-offs via reasoned, evidence-based reflection rather than intuition or random discretion." Therefore, such documents charge developers with *ex ante* fairness control, assigning them unusual compliance responsibilities and shifting the duty of defining fairness from society to companies.

We argue that the best way for developers and their employers to approach such ethical dilemmas and embed social values into algorithms is to follow a two-step approach. First, identify the algorithm's application domain—for example, it might be education, employment, or criminal sentencing. In this vein, the IEEE's Ethically Aligned Design says that embedding social values in the algorithm "requires a clear delineation of the community in which the autonomous and intelligent systems are to be deployed."[62]

---

[59] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai, 'Man is to computer programmer as woman is to homemaker? debiasing word embeddings.' In Advances in neural information processing systems (2016) 4349-4357.
[60] ibid.
[61] European Commission, 'The Ethics Guidelines for Trustworthy Artificial Intelligence' (April 2019) 13 https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Diversity
[62] IEEE (n 18) 36 Op. cit.

Second, developers and their companies should analyze case law in the relevant domain to identify the dominant fairness paradigm—anticlassification or antisubordination—and then, in accordance with the guidelines described in the previous section, choose a fairness definition compatible with that paradigm. For example, in case law addressing US university admission rules, antisubordination seems to be the dominant approach. Thus, a reasonable approach may be to create an algorithm that satisfies the equality-of-opportunity definition of fairness, because it comports with antisubordination. Another example is recidivism-risk assessment in criminal sentencing. In *The State of Wisconsin v. Eric Loomis,* the court followed the anticlassification approach in claiming that the COMPAS algorithm—which is based on the calibration definition of fairness—is legal. Hence, in this domain, satisfying the calibration definition may be sufficient for ML.

Despite these recommendations, we suggest that as long as society lacks a unified approach to fairness, developers will never be free from criticism, as ML algorithms that are fair by one definition are always unfair by other definitions. The analysis of case-law approaches to fairness for the relevant domains minimizes, but does not eliminate, the risk that developers will be accused of creating an unfair algorithm.

As the adoption of ML expands, and as the number of cases in which courts rule ML algorithms unfair increases, society will face a growing need to choose a unified notion of fairness. In other words, the large-scale use of ML is likely to drive society from abstract rules which allow a flexible interpretation of fairness to a more positivistic approach in which legal documents precisely define fairness and leave little or no room for discretion.

Formalizing a definition of fairness is a difficult task which involves morality, ethics and culture, however we claim that it is the only way to make the legal framework transparent and to avoid unproductive disputes about the fairness of AI. Tackling this challenge should enable an objective evaluation of the fairness of AI systems which make decisions affecting people's lives. This radical shift from the status quo of an abstract fairness principle to a new, precise, and positivistic definition is yet to come.

Contact details:

Ekaterina V. Semenova
National Research University Higher School of Economics. BRICS Competition Centre.
E-mail: evsemenova@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.