EXPLAINABLE FAIRNESS IN REGULATORY ALGORITHMIC AUDITING

*Cathy O'Neil, Holli Sargeant, Jacob Appel**

WEST VIRGINIA LAW REVIEW, forthcoming 2024

Abstract. How does a regulator know if an algorithm is compliant with existing anti-discrimination law? This is an urgent question, as algorithmic decision-making tools play an increasingly significant role in the lives of humans, especially at critical junctures such as getting into college, getting a job, getting a mortgage, housing, or insurance. In each of these regulated situations, moreover, the legal meaning of unlawful discrimination is different and context dependent. Regulators lack consensus on how to audit algorithms for discrimination. Recent legal precedent provides some clarity for review and provides the basis of the framework for algorithmic auditing outlined in this article. This article provides a review of precedent, a novel framework which explicitly decouples technical data science questions from legal and regulatory questions, an exploration of the framework's relationship to disparate impact. The framework promotes algorithmic accountability and transparency by focusing on explainability to regulators and the public. Through case studies in student lending and insurance, we demonstrate operationalizing audits to enforce fairness standards. Our goal is an adaptable, robust framework to guide anti-discrimination algorithm auditing until legislative interventions emerge. As an ancillary benefit, this framework is robust, easily explainable, and implementable with immediate impacts to many public and private stakeholders.

Keywords: Algorithmic Auditing, Algorithmic Bias, Anti-discrimination law, Regulatory Enforcement

* Cathy O'Neil is the CEO of ORCAA, an algorithmic auditing company, and a member of the Public Interest Tech Lab at Harvard Kennedy School. Holli Sargeant is a Ph.D. Candidate at the Faculty of Law, University of Cambridge; Visiting Researcher at Harvard Law School; and Affiliate at the Berkman Klein Center for Internet & Society. Jacob Appel is the Chief Strategist at ORCAA. We thank Deborah Hellman, Daniel Schwarcz, Christopher Bavitz, and Kevin Klyman for helpful comments and conversations on earlier drafts.

I. Introduction

The integration of algorithms into decision-making processes has increasingly become ubiquitous. Given the known flaws in human decision-making, algorithmic decision-making holds the promise of greater efficiency, consistency, and improved insights. As a result, private companies and government agencies alike have rapidly adopted these tools in an effort to aid under-resourced and overwhelmed staff. Yet, as numerous examples demonstrate, the risk of algorithmic bias and discrimination is now widely recognized.[1] This raises an urgent need for practical frameworks to guide regulators in auditing algorithms to achieve lawmakers' demands for unbiased outcomes. Defining what constitutes an *unbiased* algorithm and implementing regulatory rules to ensure compliance poses a challenge. Criteria to evaluate an algorithm's fairness, how regulators can monitor compliance, potential punitive actions, and corporate compliance represent areas of potential debate.

A common use case emerges. A large company seeks to fill an open position from a substantial number of applicants. Use of an algorithm presents a substantial time savings in addition to a clear use of logic if ever audited. The conclusion appears to be self-evident, however, in their current form, recruitment algorithms can be biased and perpetuate inequality.[2] This issue is not unique to hiring; lenders also face the challenge of evaluating the creditworthiness of applicants based on an overwhelming amount and variety of information, while insurers face a similar situation when inferring the appropriate terms of a claim. There are considerable legal uncertainties and risks associated with deploying either in-house or third-party tools in these decision-making processes. Such algorithms are widespread and important, but they can also be opaque, mysterious, and ultimately harmful. In particular, algorithms that take human traits and quantify them regularly

---

[1] See Anthony Kelly, *A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020*, 47 BRIT. EDUC. RES. J. 725 (2021) (explaining the widely criticized use of calculated grades systems in the UK and their unfair outcomes for students); Ziad Obermeyer et al., *Dissecting racial bias in an algorithm used to manage the health of populations*, 366 SCIENCE 447 (2019) (showing evidence of racial bias in the widely used algorithm from health services company, Optum); Taylor Telford, *Apple Card Algorithm Sparkes Gender Bias Inquiry*, WASHINGTON POST (2019), https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/ (discussing allegations of gender bias in the algorithm used by Goldman Sachs to determine credit limits for the Apple Card); Jeffrey Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, REUTERS (Oct. 10, 2018) https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (reporting on a recruiting tool developed by Amazon that used artificial intelligence but was scrapped due to concerns about bias against women); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016) (providing many example, including a teacher, Sarah Wysocki, that was unfairly fired by algorithm); Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (analyzing the COMPAS algorithm, which is used to predict recidivism in the criminal justice system, and found evidence of racial bias).

[2] Khari Johnson, *Feds Warn Employers Against Discriminatory Hiring Algorithms*, WIRED (May 16, 2022), https://www.wired.com/story/ai-hiring-bias-doj-eecc-guidance/; Miranda Bogen, *All the Ways Hiring Algorithms Can Introduce Bias*, HARVARD BUSINESS REVIEW (May 6, 2019), https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias; Dastin, *supra* note 2.

result in damaging effects and the perpetuation of bias–they can even be Weapons of Math Destruction, or WMDs.[3]

Assuming businesses do in fact fully understand the harms of bias, there are still no incentives to audit or change algorithms. Reputational benefits companies gain from demonstrating the fairness of implemented AI systems generally do not outweigh the risk that self-auditing and disclosure of bias may pose. In fact, few companies have faced legal consequences, even when algorithms have been publicly identified as discriminatory.[4] Those companies that do incur penalties do not appear to face proportionate consequences.[5] The proprietary nature of algorithms, the cost and complexity of conducting audits, and a lack of legal backstops contribute to this incentive impasse.

Many lawmakers, regulators, and scholars propose algorithmic auditing, based on disparate impact doctrine, as a solution to these problems.[6] Understandably, private companies have very little guidance on how to adopt these tools without risking discrimination liability.

The lack of regulatory clarity equally contributes to the lack of incentives. It is in the best interest of companies to interpret new laws in their narrowest scope. Regardless of potential moral hazard, even where there is a will to change and avoid discriminatory algorithms, a lack of tangible instruction and guidance does not encourage such change. In addition, in practice it is difficult to rigorously test algorithms for discrimination because most scoring systems are shrouded in secrecy.[7] Most algorithmic designers[8] refuse to reveal the method and logic of predictive systems, meaning algorithms are zealously guarded trade secrets.[9]

This Article proposes a novel "Explainable Fairness" framework to provide a practical guide to regulators in auditing algorithms for compliance. The proposal builds on relevant legal precedents, while translating computer science literature into actionable guidelines. A key focus is promoting transparency and explainability to address the opacity of commercial algorithms that often hinders oversight.

---

[3] O'NEIL, *supra* note 2.
[4] Alex Engler, Independent Auditors Are Struggling to Hold AI Companies Accountable, Fast Company (Jan. 26, 2021), https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue; Alex Engler, Auditing Employment Algorithms for Discrimination, Brookings (Mar. 12, 2021), https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination/.
[5] Engler, *supra* note 5.
[6] *See infra* Section II.B.
[7] *See* FRANK PASQUALE, THE BLACK BOX SOCIETY (2019); Danielle Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. LAW REV. 1, 5 (2014).
[8] In this Article, the "regulated entity" may be either (or both) the algorithm maker or the company deploying the algorithm. The distinction is not pursued in detail because it may be the same entity, and some laws (especially new algorithmic audit laws, *infra* Section II.B.) require either the algorithmic vendor or user to comply. This question more broadly should be answered by each regulator based on the respective laws and industry compliance.
[9] *See* PASQUALE, *supra* note 8; Citron and Pasquale, *supra* note 8.

In recent years, thousands of papers have been written on algorithmic fairness.[10] However, there is a dearth of pragmatic and meaningful proposals for law reform. In particular, the computer science-led discipline struggles to translate and apply existing laws and evolving regulations to real-world projects. Without clarity on how legal doctrine should be applied to algorithmic audits, regulators will struggle with enforcement and individuals may not benefit from these new protections. This article fills the gap by setting forth a robust framework regulators can employ to audit algorithms, set enforceable fairness standards, and identify discrimination - providing an interim solution until legislative interventions emerge and an iterative solution that can evolve with changing regulation.

Our approach will help guide sufficiently resourced regulators in the current milieu of rapid legal and technological development. It will help define fairness in a specific, and therefore normatively relevant context, set thresholds of enforcement acceptability, and help regulators implement disparate impact audits.

Part II considers existing literature on algorithmic fairness and discrimination and examines recent legal developments shaping the current regulatory landscape. Part III introduces our proposal for explainable fairness. Part IV offers explanations of how this framework can be operationalized and applied through two use cases – student lending and disability insurance. In this section, the Article engages with the most pressing challenges related to algorithmic auditing and highlight the need for future work to understand the implications of this framework for regulators, companies, and society as a whole. Finally, Part V concludes and proposes opportunities for implementation of Explainable Fairness.

II. Background on Algorithmic (Un)Fairness

A. Algorithmic fairness literature

Algorithmic fairness is an area of literature that explores the technical, ethical, and legal concerns with algorithmic decision-making. The literature on algorithmic fairness and accountability has

[10] In subject-matter conferences alone, there have been 773 papers accepted since 2018, *see FAccT Conference*, ACM, https://facctconference.org/ (last visited Apr. 11, 2023); *Artificial Intelligence, Ethics, and Society Conference*, AAAI/ACM, https://www.aies-conference.com/ (last visited Apr. 11, 2023); Other traditional machine learning conferences have also expanded to include fairness, accountability and transparency tracks, *see NeurIPS Conference*, NEURAL INFORMATION PROCESSING SYSTEMS FOUNDATION, https://nips.cc/ (last visited Apr. 11, 2023); *ICML Conference*, INTERNATIONAL CONFERENCE ON MACHINE LEARNING, https://icml.cc/ (last visited Apr. 11, 2023).

exploded alongside the rapid evolution of algorithmic tools in recent years.[11] Since 2014, scholars from a range of disciplines, although primarily computer scientists, have convened at the annual Fairness, Accountability, and Transparency (FAccT) Conference to discuss the latest in how machine learning (ML) algorithms can be made more accountable in various circumstances and contexts.[12]

Despite this rapidly evolving field, the definition of "algorithmic bias" is still a hotly contested topic in ML literature, particularly as different approaches seek to deal with different types of harm.[13] Many scholars have made valuable contributions seeking to define what fairness means with respect to algorithmic bias.[14] Each definition has broadly similar intent, some being the definition of when algorithms systematically perform worse, or penalize, certain groups of people.[15] Much of this work is spearheaded by the ML science community, which has proposed measuring fairness using statistical metrics that attempt to measure disparity between different individuals and groups.

However, the literature suffers from being either highly technical (if it originates in the academic computer science community) or being highly critical without clear suggestions for technical support (if it originates in the legal or sociological community). A cohesive understanding of these two disparate fields in the literature is difficult to achieve. For instance, the 2019 FAccT Conference tutorial that addressed "21 fairness definitions and their politics",[16] serve to make the task bewildering rather than straightforward. Generally, it has been unhelpful to describe different statistical measures as different conceptions of fairness.[17]

---

[11] *Id.*
[12] *Fairness, Accountability, and Transparency in Machine Learning Conference*, FAT ML, https://www.fatml.org/schedule/2014 (last visited Apr. 11, 2023).
[13] Solon Barocas et al., *Introduction*, *in* FAIRNESS AND MACHINE LEARNING: LIMITATIONS AND OPPORTUNITIES (2022) (explaining allocative and representational harms).
[14] *See e.g.*, Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy* 149 (Association for Computing Machinery Mar. 2021); Alice Xiang & Inioluwa Deborah Raji, *On the Legal Compatibility of Fairness Definitions*, Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (Association for Computing Machinery Sept. 2019); Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, Proceedings of the International Workshop on Software Fairness 1 (Association for Computing Machinery May 2018); Jon Kleinberg et al., *Algorithmic Fairness*, 108 AEA PAPERS AND PROCEEDINGS 22 (May 2018); Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 797 (Association for Computing Machinery Aug. 2017).
[15] *See e.g.*, Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 SOC. METHODS & RES. 3 (2021); Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. LAW REV. 649 (2021).
[16] Arvind Narayanan, *Tutorial: 21 Fairness Definition and Their Politics*, in CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (ACM 2018) https://fairmlbook.org/tutorial2.html.
[17] Berk et al., *supra* note 16 (describing six statistical measures of fairness); Verma & Rubin, *supra* note 15 (exploring definitions for 20 different statistical measures related to algorithmic fairness).

So, while this research enriches the academic conversation, it provides little practical guidance for those who must operationalize fairness in specific algorithmic contexts and perhaps less for lawmakers or regulators. The expectation, implicitly arising from the technical literature, that civil servants become professional mathematicians to choose the most suitable fairness metric is not implementable. It yields an alternative where these crucial decisions are relegated to computer scientists, bypassing necessary legal oversight.

Several legal scholars have made strides to translate algorithmic fairness literature to legal audiences.[18] In particular, legal scholars have identified the tensions between statistical measures of disparity and current legal frameworks for unfairness, primarily in anti-discrimination law.[19] Many fairness metrics aim to evaluate an algorithm with respect to its performance across subgroups with certain attributes, usually borrowing legally protected characteristics from anti-discrimination law.[20] Therefore, comparison between classification and effects on individuals based on protected characteristics reveals the overlap in this field.[21] Despite best intentions to map statistical measures of disparity to law, many legal scholars arrive at the conclusion that there is a separate exercise between conceptions of algorithmic fairness and algorithmic *unlawful* discrimination.[22]

B. Algorithmic Discrimination

The idea that algorithms can unjustifiably discriminate is not new and is the focus of considerable legal research on algorithms.[23] The focus in recent literature has been how algorithmic bias – the

---

[18] *See, e.g.*, Jeremias Adams-Prassl et al., *Directly Discriminatory Algorithms*, 86 MOD. L. REV. 144 (2023); Xiang, *supra* note 2; Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. LAW REV. 811 (2020); Talia Gillis & Jann Spiess, *Big Data and Discrimination*, 86 UNIV. CHIC. LAW REV. 459 (2019); Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113 (2018).

[19] Talia Gillis, *The Input Fallacy*, 106 MINN. L. REV. 1175 (2022); Xiang, *supra* note 16.

[20] Shira Mitchell et al., *Algorithmic Fairness: Choices, Assumptions, and Definitions*, 8 ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION 141 (2021); Corbett-Davies et al., *supra* note 5; Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS Oct. 2016).

[21] See, e.g., Xiang & Raji, *supra* note 15; Thomas B Nachbar, *Algorithmic Fairness, Algorithmic Discrimination*, 48 FLA. ST. U. L. REV. 509 (2021); Kleinberg et al., *Algorithmic Fairness*, *supra* note 15.

[22] Nachbar, *supra* note 22, at 524:

> "And here lies a fundamental difference between fairness and law. Unlike fairness, law cannot exist as an inherently contested concept because the foundation of law is agreement among society (at least generally) as to its content. Law—the judicially enforceable rules by which society operates—reflects a settlement of inherently unresolvable conflicts regarding fairness, or right, or justice."

[23] *See, e.g.*, Nachbar, *supra* note 22; Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. LAW REV. 671 (2016); Jason R Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803 (2020); Hellman, *supra* note 8; Pauline Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE. (Jan. 2017), https://scholarship.law.upenn.edu/penn_law_review_online/vol166/iss1/10; Joshua Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017); Finn Lattimore, Simon O'Callaghan, Zoe Paleologos, Alistair Reid, Edward Santow, Holli Sargeant & Andrew Thomsen, Using Artificial Intelligence to Make Decisions: Addressing the Problem of Algorithmic Bias, Australian Human Rights Commission (2020),

less favorable treatment or penalization of individuals by an algorithm – fits within current anti-discrimination doctrine. It is not a question of when, but rather a question of how.[24]

1. Disparate treatment and disparate impact in algorithmic discrimination

Directly discriminatory algorithms have often been dismissed because disparate treatment doctrine liability focuses on formal classifications and intentional discrimination.[25] That is not to say there has not been important contributions on cases of disparate treatment.[26] It is important to clarify that the "concept of disparate treatment is more elusive than is often recognized."[27]

Literature on algorithmic discrimination has still proceeded largely on the basis of disparate impact doctrine. Disparate impact protects against a policy or practice that is facially neutral but has a disproportionately adverse impact on protected classes.[28] The theory of disparate impact was initially a product of judicial interpretation,[29] and was eventually strengthened through codification.[30] Disparate impact liability arises across various sectors–including employment,[31] housing,[32] and lending.[33]

---

https://humanrights.gov.au/sites/default/files/document/publication/ahrc_technical_paper_algorithmic_bias_2020.pdf.

[24] Nachbar, *supra* note 22, at 543 (emphasizing there is no question anti-discrimination law will be applied to algorithms, but rather "how discrimination law will have to adapt to computational decisionmaking and how computational decision-making will have to adapt to discrimination law.").

[25] Barocas & Selbst, *supra* note 24, at 672 (suggesting that "the best doctrinal hope for data mining's victims would seem to lie in disparate impact doctrine").

[26] *See, e.g.*, Pauline Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (Feb. 2017); Crystal Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291 (2020); Hellman, *supra* note 19; Bent, *supra* note 24.

[27] Hellman, *supra* note 19, at 819.

[28] Griggs v. Duke Power Co., 401 U.S. 424 (1971).

[29] *Id.*

[30] Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071 (codified in scattered sections of 42 U.S.C.).

[31] Civil Rights Act of 1964, Pub. L. No. 88-352, tit. VII, §§ 701-716, 78 Stat. 253 (codified as amended in scattered sections of 42 U.S.C.).

[32] Texas Department of Housing and Community Affairs v. The Inclusive Communities Project, Inc., 576 U.S. 519 (2015) (SCOTUS held that the Fair Housing Act of 1968 creates a cause of action for disparate impact).

[33] Equal Credit Opportunity Act (Title VII of the Consumer Credit Protection Act), 15 U.S.C. § 1691(a) (1974); Equal Credit Opportunity Act (Regulation B), 12 CFR § 1002 (2012); Cf. Comment for 1002.6—Rules Concerning Evaluation of Applications, CONSUMER FINANCIAL PROTECTION BUREAU, https://www.consumerfinance.gov/rules-policy/regulations/1002/interp-6/ (last visited Apr 26, 2023) (CFPB interpreting ECOA to include disparate impact liability, consistent with judgements from circuit courts and district courts); Peter N. Cubita & Michelle Hartmann, *The ECOA Discrimination Proscription and Disparate Impact—Interpreting the Meaning of the Words That Actually Are There*, 61 BUS. LAW. 829 (2006), https://www.jstor.org/stable/40688368 (arguing that Congress did not intend for the disparate impact method of proving discrimination claims to apply in ECOA litigation because neither the ECOA's statutory discrimination proscription nor its legislative history supports a finding that the Act prohibits facially neutral practices that disparately affect protected class members).

In a disparate impact case, the investigation focuses on the consequences of the respondent's practice rather than intent.[34] It follows that the majority of the new and proposed auditing laws require an assessment of the tool's disparate impact and do not address disparate treatment.[35]

Generally, disparate impact liability is made out by three important steps: (1) the plaintiff's burden of proving the defendant uses a practice that causes disparate impact, (2) the defendant's burden of proving any reasonable justification through business necessity for such practices, and (3) whether any available alternative practices could have been used with less discriminatory impact.[36]

2. Establishing disparate impact in algorithmic systems

First, to establish a prima facie case for disparate impact, a plaintiff must identify a specific policy or practice that causes disparate impact on the basis of a protected attribute.[37] Initially, identifying a particular policy or practice that has created the disparate impact is an integral part of the prima facie case in a disparate impact suit.[38] Identifying a particular practice can be a high barrier for some plaintiffs, and many individuals are unaware of the use of algorithmic practices, but on the other hand, where known by a complaining party, "algorithmic selection procedures give potential plaintiffs an obvious target."[39]

Once a specific practice is identified, i.e., use of an algorithmic system, a complaining party must show evidence of a disparity. Plaintiffs must show there is a pattern based on protected attributes that is "significantly different" from the pool of individuals subject to the policy or practice.[40] It is not entirely clear whether "significantly different" refers to a formal statistical disparity or if it is

---

[34] Lau v. Nichols, 414 U.S. 563, 568 (1974); Barocas and Selbst, *supra* note 24, at 701 (explaining "[w]here there is no discriminatory intent, disparate impact doctrine should be better suited to finding liability for discrimination in data mining").
[35] *See* N.Y.C., N.Y., Admin. Code § 20-870 (2021) (Local Law 144 to amend the administrative code of the city of New York, in relation to automated employment decision tools); D.C., Council B. 240558, 24th Council, (D.C. 2021) (Stop Discrimination by Algorithms Act of 2021).
[36] Albermarle Paper Co. v. Moody, 422 U.S. 405 (1975); The "alternative employment practice" test was not always treated as a separate test, *see* Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989) however, the approach was codified in the 1991 Act, 42 U.S.C. § 2000e-2(k)(1)(A).
[37] For example, in respect of employment, the plaintiff's initial burden is to prove that the "respondent uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex or national origin", *see* 42 U.S.C. § 2000e-2(k)(1)(a)(i).
[38] Watson v. Ford Worth Bank & Trust, 487 U.S. 977, 994 (1988) ("the plaintiff's burden in establishing a prima facie case goes beyond the need to show that there are statistical disparities in the employer's workforce. The plaintiff must begin by identifying the specific employment practice that is challenged.... Especially in cases where an employer combines subjective criteria with the use of more rigid standardized rules or tests").
[39] Matthew U. Scherer, Allan G. King & Marko J. Mrkonich, *Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms*, 71 S. C. L. REV. 449, 495 (2020); Colin Clemente Jones, *Systematizing Discrimination: AI Vendors & Title VII Enforcement Comment*, 171 U. PA. L. REV. 236, 256 (2022–2023).
[40] Albermarle Paper Co. v. Moody, *supra* note 37, at 425; Ricci v. DeStefano, 557 U.S. 557, 857 (2009) (explaining that a prima facie case of disparate impact requires showing a statistically significant disparity "and nothing more.").

meant in some more colloquial sense.[41] In any event, as discussed in greater detail below, algorithmic systems at scale can render even slight differences in selection rates statistically significant,[42] so it will be important to consider the different statistical approaches courts have considered in deciding what a statistically significant difference means.

Second, the disparate impact case is progressed if a complaining party establishes this prima facie case to shift the burden to the respondent to prove that such practice "consistent with business necessity."[43] The Court's discussion and interpretation of the business necessity defense has varied.[44] The most recent comment from the Court in Inclusive Communities set a low threshold for what constituted business necessity: "private policies are not contrary to the disparate-impact requirement unless they are 'artificial, arbitrary, and unnecessary barriers'."[45] Determining what is legitimate target criteria to meet business necessity is one, challenging aspect of algorithmic disparate impact. This Article discusses how legitimate factors could be identified and justified in algorithmic systems in detail below.

Third, if the respondent demonstrates a legitimate business necessity, a plaintiff must show "that there is 'an available alternative . . . practice that has less disparate impact and serves the [entity's] legitimate needs.'"[46] In *Albermarle*, this included "other tests or selection devices, without a similarly undesirable racial effect."[47] The plaintiff's burden of proving the existence of an alternative, less discriminatory, and equally effective practice that serves legitimate business needs sets an impossible threshold for plaintiffs because it is incompatible with algorithmic decision-making. As most algorithms are black boxes, it will be almost impossible for a plaintiff to offer any alternative.[48] Businesses that use third-party models pose an unsolvable problem as these models cannot be manipulated or interrogated. Employing a businesses' own model removes this obstacle but does not guarantee a viable alternative.

Perhaps a plaintiff could design their own algorithm to attempt to reveal substantially equivalent alternatives. The technical burden on a plaintiff and their legal team to undertake such an exercise would be considerable and would exceed the intentions of the law and impose an undue burden on the plaintiff.[49] Any analysis of alternative models, datasets or model specifics would be difficult to compare without knowing the underlying construct of relevant models. Consider an algorithm

---

[41] Albermarle Paper Co. v. Moody, *supra* note 37, at 425; Scherer, King, and Mrkonich, *supra* note 40, at 464.
[42] Scherer, King, and Mrkonich, *supra* note 40, at 464.
[43] 42 U.S.C. § 2000e-2(k)(1)(a)(i).
[44] *See* Griggs v. Duke Power Co., 401 U.S. 424 (1971); Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 659 (1989); Albemarle Paper Co. v Moody, 422 U.S. 405 (1975); Texas Department of Housing and Community Affairs v. The Inclusive Communities Project, Inc., 576 U.S. 519 (2015).
[45] *Inclusive Communities*, *supra* note 33, at 537; citing Griggs v. Duke Power Co., *supra* note 29, at 431.
[46] *Inclusive Communities*, *supra* note 33, at 537; citing Ricci v. DeStefano, *supra* note 41, at 578.
[47] Albermarle Paper Co. v. Moody, *supra* note 37, at 425.
[48] Michael Selmi, *Algorithms, Discrimination and the Law*, 82 OHIO ST. L.J. 406, 644 (2020); Barocas and Selbst, *supra* note 24, at 710.
[49] *Id*. at 644.

that would have improved predictive accuracy if a business collected the maximum amount of data. Without legal standards for predictive performance or accuracy, it is difficult to design a true alternative to additional data collection, different datasets, or label selections.[50]

Further, achieving less discriminatory impact by using protected characteristics in the model "presents a tension between the input-focused 'disparate treatment' and the outcome focused 'disparate impact' doctrines".[51] Indeed, including these attributes may allow for mitigation of harm, for example, by being able to directly decorrelate the outputs by race.[52]

Given the plaintiff's almost insurmountable barriers in providing an alternative, less discriminatory practice,[53] there should be opportunities for relief to discriminated individuals in algorithmic contexts that do not rely exclusively on disparate impact litigation.

3. Actionable algorithmic discrimination

The disparate impact doctrine will not always apply to instances of discriminatory algorithmic decision-making because it is limited to certain regulated areas. Even recently, the Consumer Financial Protection Bureau (CFPB) announced it would target discrimination more broadly through the agency's regulatory authority over unfair, deceptive, or abusive acts or practices.[54] According to the CFPB, this means the agency will regulate unfair discrimination across every aspect of every financial services provider over which the Bureau has jurisdiction.[55]

Insurance discrimination law, in particular, is a complex area of law and unique from civil rights anti-discrimination laws.[56] While some insurance, especially home insurance or types of credit insurance might fall under federal civil rights legislation, most insurance regulation is delegated

[50] *See* Virginia Foggo & John Villasenor, *Algorithms, Housing Discrimination, and the New Disparate Impact Rule*, 22 COLUM. SCI. & TECH. L. REV. 1 (2020–2021).
[51] Gillis, *supra* note 20, at 49; Gillis and Spiess, *supra* note 19, at 472.
[52] Gillis, *supra* note 20, at 49; Gillis and Spiess, *supra* note 19, at 472.
[53] Selmi, *supra* note 49, at 644.
[54] The updated UDAAP Examination Manual urges examiners to consider whether a supervised entity has "a process to take prompt corrective action if the decision-making process it uses produce deficiencies or discriminatory results", and ensure that employees and third party service providers "refrain from engaging in servicing or collection practices that lead to differential treatment or disproportionately adverse impacts on a discriminatory basis", Consumer Financial Protection Bureau, *Unfair, Deceptive, or Abusive Acts or Practices (UDAAPs) Examination Manual*, (2022), https://files.consumerfinance.gov/f/documents/cfpb_unfair-deceptive-abusive-acts-practices-udaaps_procedures.pdf (last visited May 28, 2023); CFPB Targets Unfair Discrimination in Consumer Finance, CONSUMER FINANCIAL PROTECTION BUREAU (2022), https://www.consumerfinance.gov/about-us/newsroom/cfpb-targets-unfair-discrimination-in-consumer-finance/ (last visited May 24, 2023).
[55] CFPB Targets Unfair Discrimination in Consumer Finance, *supra* note 55.
[56] *See generally* KENNETH S. ABRAHAM & DANIEL SCHWARCZ, INSURANCE LAW AND REGULATION (6th ed. Foundation Press 2015).

to the states.[57] In many cases, insurers are allowed to differentiate among individuals based on actuarially sound principles, such as the use of actuarial data to determine risk. Such practices are not subject to disparate impact scrutiny.[58] Insurers are generally prohibited from treating individuals in the same risk class differently amounting to 'unfair discrimination'.[59] Some states also provide legislation that prohibit insurance decisions based on race,[60] or based solely on sensitive attributes.[61] Given that algorithmic decision-making leading to disparate impact that is actuarially justified is permissible in insurance, while under a heavy justificatory burden in other contexts, the scope of algorithmic discrimination is broad and complex.

4. Bridging the legal-technical divide

Legal interpretations concerning algorithmic discrimination often reside in a state of ambiguity, with experts navigating through best estimations. For example, the Ricci decision,[62] has been subject to academic discussions on the implications of the Ricci decision in an algorithmic context.[63] The point of debate here is whether the Ricci decision applies to the realm of algorithms

---

[57] Kudakwashe F. Chibanda, *Defining Discrimination in Insurance*, CAS RESEARCH PAPER SERIES ON RACE AND INSURANCE PRICING (Casualty Actuarial Society 2022); *see also* Daniel Schwarcz, *Towards a Civil Rights Approach to Insurance Anti-Discrimination Law*, 69 DEPAUL L. REV. 657 (2020).

[58] NATIONAL ASSOCIATION OF MUTUAL INSURANCE COMPANIES, THE LEGAL THEORY OF DISPARATE IMPACT DOES NOT APPLY TO THE REGULATION OF CREDIT-BASED INSURANCE SCORING (July 7, 2004).

[59] Casualty Actuarial Society, *Statement of Principles Regarding Property and Casualty Insurance Ratemaking*, https://www.casact.org/statement-principles-regarding-property-and-casualty-insurance-ratemaking (last visited Apr. 26, 2023) (Principle 4: "A rate is reasonable and not excessive, inadequate, or unfairly discriminatory if it is an actuarially sound estimate of the expected value of all future costs associated with an individual risk transfer."); a few states also have laws that provide that rates based on certain factors, such as race and sex are, by definition, unfairly discriminatory. *See e.g.*, TENN. CODE ANN. § 56-5-302 (LEXISNEXIS THROUGH 2023 REG. SESS.); WYO. STAT. ANN. § 26-14-103 (LEXISNEXIS THROUGH 2023 GEN. SESS.).

[60] *See, e.g.*, CONN. GEN. STAT. § 38A-358 (LEXISNEXIS THROUGH 2023 REG. SESS.); TEX. INS. CODE ANN. § 544.002 (LEXISNEXIS THROUGH 2021 REG. SESS.); N.C. GEN. STAT. § 58-3-25 (LEXISNEXIS THROUGH SESSION LAWS 2023-12).

[61] *See, e.g.,* ARIZ. REV. STAT. ANN. § 20-1631 (LEXISNEXIS THROUGH 56TH LEG. 1ST REG. SESS. 2023); ARK. CODE ANN. § 23-66-206 (LEXISNEXIS THROUGH APRIL 6, 2023); COLO. REV. STAT. § 10-4-626 (LEXISNEXIS THROUGH 2023 REG. SESS.); FLA. STAT. ANN. § 634.282 (LEXISNEXIS THROUGH 2023 B SPEC. SESS.); HAW. REV. STAT. § 431:10C-111 (LEXISNEXIS THROUGH 2030 LEGIS. SESS.); IOWA CODE § 515F.6 (LEXISNEXIS THROUGH 2022 REG. SESS.); KY. REV. STAT. ANN. § 304.20-040 (LEXISNEXIS THROUGH 2023); LA. REV. STAT. ANN. § 22:35 (LEXISNEXIS THROUGH 2023); MO. REV. STAT. § 379.114 (LEXISNEXIS THROUGH 2022); NEV. REV. STAT. ANN. § 687B.390 (LEXISNEXIS THROUGH 82ND REG. SESS. 2023); N.H. REV. STAT. ANN. § 417-A:3 (LEXISNEXIS THROUGH 2023 REG. SESS.); N.Y. INS. LAW § 2606(B) (MCKINNEY 2023); OHIO REV. CODE ANN. § 3937.39 (LEXISNEXIS THROUGH 2023-2024); S.D. CODIFIED LAWS § 58-11-55 (LEXISNEXIS THROUGH 2023 REG. SESS.); VA. CODE ANN. § 38.2-2213 (LEXISNEXIS THROUGH 2023 SESS.).

[62] Ricci v. DeStefano, *supra* note 41 (grappling with a scenario where an action, seemingly favoring one race, was annulled due to its potential discriminatory consequences. However, the Ricci case involved specific individuals who suffered due to a race-based decision, resulting in a policy reversal. This situation contrasts significantly with algorithmic policies, which are largely predicated on broad statistical trends rather than individual instances.)

[63] *See* Ricci v. DeStefano, *supra* note 41; Kroll et al., *supra* note 24 ("If an agency runs an algorithm that has a disparate impact, correcting those results after the fact will trigger the same kind of analysis as New Haven's rejection of its firefighter test results"); Barocas and Selbst, *supra* note 24 (arguing that Ricci prohibits an employer from making changes to an algorithm after seeing that it will have a disparate impact on racial minorities).

at all. Legal scholars suggest that the Ricci precedent does not quite fit. Their consensus is that if a computational model is analyzed and identified as having a potential for discrimination, and amendments are made to redress this issue, it should not present any legal complications.[64] However, this debate exemplifies the risk that the misinterpretation of legal concepts could halt important mitigations trying to remove algorithmic bias. For example, certain approaches to algorithmic fairness require assumptions that are intrinsically value-laden and legal exercises.[65] Assessing conditional statistical parity or individual fairness notions requires the identification of legitimate or relevant features, not just differences in means.[66] For that matter, the "law's treatment of explicit racial classifications is more complex and nuanced than scholars writing about algorithms have recognized thus far."[67] Even technical focus on statistical disparity, including the omnipresence of the four-fifths rule,[68] is helpful but does not fully capture the scope of disparate impact doctrine.[69] Technical strategies are necessary but have limitations with respect to discrimination as they cannot be perfectly consistent with what the law intends.

Pauline Kim writes: "When it comes to the goals of nondiscrimination, however, purely technical strategies cannot guarantee that automated decision processes will be free of bias. These processes may systematically disadvantage protected groups as a result of social processes that lie outside the code".[70]

C. Emerging Algorithmic Legal Developments

In recent years, American policymakers have become increasingly cognizant of the hazards posed by algorithms and are now seeking to mandate unbiased algorithms. In order to ground the discussions in this Article, it is helpful to begin with an overview of emerging legal developments.

---

[64] Hellman, *supra* note 19; citing Kim, *supra* note 24, at 191 (arguing that Kroll misreads Ricci and that that case "narrowly addressed a situation in which an employer took an adverse action against identifiable individuals based on race, while still permitting the revision of algorithms prospectively to remove bias"); Kim, *supra* note 27, at 869.
[65] Kim, *supra* note 24; Hellman, *supra* note 19.
[66] *See* Kim, *supra* note 24; Kroll et al., *supra* note 24; Cynthia Dwork et al., Fairness Through Awareness, in PROCEEDINGS OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE 214 (2012), https://doi.org/10.1145/2090236.2090255.
[67] Hellman, supra note 19, at 855.
[68] *See, e.g.*, Christo Wilson et al., Building and Auditing Fair Algorithms: A Case Study in Candidate Screening, in PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 666 (2021), https://doi.org/10.1145/3442188.3445928 (last visited Feb 27, 2023).
[69] U.S. Equal Employment Opportunity Commission, *Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964*, (2023), https://www.eeoc.gov/select-issues-assessing-adverse-impact-software-algorithms-and-artificial-intelligence-used (last visited May 28, 2023).
[70] Kim, *supra* note 24, at 202.

1. Legislative changes and proposals

At a federal level, there has been a discernable interest in legislating algorithmic accountability, although progress to date has been limited. The Algorithmic Accountability Act was first proposed in 2019, and despite being reintroduced in the 2022 Congress, has only made modest headway.[71] Senator Ron Wyden, who proposed the bill, stated that it mandates "companies to study the algorithms they use, identify bias in these systems and fix any discrimination or bias they find."[72]

Notably, on October 4, 2022, the White House Office of Science and Technology Policy issued a Blueprint for an AI Bill of Rights.[73] The Blueprint outlines five high-level principles to safeguard the public and protect the rights of Americans.[74] The White House's Original intention was to develop "a bill of rights for an AI-powered world",[75] but the finished document was described as a "white paper" and a "framework".[76] The absence of concrete legislation has been critiqued, especially if the aim is to foster effective AI governance in the private sector.[77]

In attempts to fill the gap left by federal inaction on legislation, state and local lawmakers have advanced proposals and enacted legislation to promote algorithmic accountability.

One example of such legislation is the Stop Discrimination by Algorithms Act of 2021, which the District of Columbia aims to pass in 2023.[78] The bill would prohibit the use of algorithmic decision-making in a discriminatory manner by organizations that possess "personal information on more than 25,000 District residents" and would require corresponding notices to individuals whose personal information is used in certain algorithms to determine decisions about employment, housing, healthcare, and lending.[79] The bill would introduce auditing of algorithms for discriminatory patterns.[80] Unlike other bills, in addition to empowering the Office of the

---

[71] Algorithmic Accountability Act of 2022, S. 3572, 117th Cong. (2022).
[72] Booker, Wyden, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms, U.S. SENATOR CORY BOOKER OF NEW JERSEY (Apr. 10, 2019), https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms.
[73] THE WHITE HOUSE OFFICE OF SCIENCE AND TECHNOLOGY POLICY, Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People (Oct. 2022), https://www.whitehouse.gov/ostp/ai-bill-of-rights/.
[74] Biden-Harris Administration Announces Key Actions to Advance Tech Accountability and Protect the Rights of the American Public (Fact Sheet AI Bill of Rights), THE WHITE HOUSE (Oct. 4, 2022), https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/.
[75] Eric Lander & Alondra Nelson, *Americans Need a Bill of Rights for an AI-Powered World*, WIRED, Oct. 8, 2021, https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/.
[76] Fact Sheet AI Bill of Rights, *supra* note 75.
[77] *See* Emmie Hine & Luciano Floridi, *The Blueprint for an AI Bill of Rights: In Search of Enaction, at Risk of Inaction*, MINDS & MACHINES (2023), https://doi.org/10.1007/s11023-023-09625-1.
[78] Statement: Councilmember Robert White Shares Path Forward for Stop Discrimination by Algorithms Act, OFFICE OF COUNCILMEMBER ROBERT WHITE (Nov. 17, 2022), https://www.robertwhiteatlarge.com/statement-stop-discrim-algorithms/.
[79] D.C., Council B. 240558, 24th Council, (D.C. 2021) (Stop Discrimination by Algorithms Act of 2021).
[80] *Id.* at § 7.

Attorney General, it would introduce a private right of action for individuals to bring suit for violation of its provisions.[81]

Massachusetts also has a pending bill that proposes an "automated decision system impact assessment" to protect consumers from potential algorithmic bias.[82]

New York City recently amended its administrative code to address automated employment decision tools.[83] Companies using automated tools for hiring and promoting employees are now required to have these systems audited by an independent entity. Local Law 144 took effect on January 1, 2023, and the New York City Department of Consumer and Worker Protection will begin enforcement on July 5, 2023.[84] The New York City Department of Consumer and Worker Protection Final Rule clarifies several elements of required bias audits.[85]

Many other states have followed suit. New Jersey introduced a nearly identical bill in 2022.[86] California has a proposed bill that would similarly introduce protections against algorithmic discrimination in hiring.[87] Specifically, the proposed rules would prohibit automated decision systems that limit, express a preference for, or screen out applicants based on protected attributes or proxies for such attributes unless there is an affirmative defense for such a criterion.[88]

The insurance sector also faces new regulations regarding its use of AI. Colorado passed new laws on algorithmic discrimination in insurance that came into effect on January 1, 2023.[89] The new law prohibits unfair discrimination and holds insurers accountable for testing and reporting on how they mitigate bias and discrimination in their data and algorithms.[90] In keeping with insurance

---

[81] *Id.* at § 8.
[82] Mass., H.B. 4029, 192nd Gen. Ct., (Mass. 2021).
[83] N.Y.C., N.Y., Admin. Code § 20-870 (2021) (Local Law 144 to amend the administrative code of the city of New York, in relation to automated employment decision tools.)
[84] *DCWP AEDT Rules Virtual Public Hearing*, (Nov. 4, 2022), https://www.nyc.gov/assets/dca/downloads/pdf/about/HearingTranscript-AEDT-Rules-Virtual-Public-Hearing.pdf (explaining that enforcement was delayed from Jan. 1, 2023, due to the substantial volume of comments received in the public hearing to clarify definitions and calculations).
[85] N.Y.C. Consumer and Worker Prot., Final Rule on Local Law 144 (2023) (to be codified at Rules of City of N.Y. tit. 6, ch. 5, § 5-300).
[86] S.B. A4909, 220th Leg., (N.J. 2022) (regulates the use of automated tools in hiring decisions to minimize discrimination in employment).
[87] Cal. Fair Emp. & Hous. Council, Proposed Modifications to Employment Regulations Regarding Automated-Decision Systems (Mar. 14, 2022), https://calcivilrights.ca.gov/wp-content/uploads/sites/32/2022/03/AttachB-ModtoEmployRegAutomated-DecisionSystems.pdf (to be codified at Cal. Code Regs. tit. 2, § 11008); a bill similar to the NYC law was proposed in 2020 but stalled in committee, S.B. 1241, 2019-2020 Reg. Sess. (Cal. 2020) (discrimination in employment: employment tests and selection procedures).
[88] *Id.*
[89] COLO. REV. STAT. § 10-3-1104.9 (2022) (ADDED BY RESTRICT INSURERS' USE OF EXTERNAL CONSUMER DATA, CH. 2887, 2021 COLO. SESS. LAWS 2887 (CODIFIED AT COLO. REV. STAT. § 10-3-1104.9 (2022)) (S.B. 21-169, 73RD GEN. ASSEMB., REG. SESS. (COLO. 2021))).
[90] *Id.* (prohibiting: "(I) Unfair discrimination based on race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression in any insurance practice; and (II) The use of external

unfair discrimination tradition, there is a carve-out for life, annuity, long-term care, or disability insures to use information with a direct relationship to mortality or longevity risk based on actuarially sound principles.[91] Even in such a case, the new law requires insurers to disclose if the algorithm uses external consumer data and information and the insurance commissioner may investigate any use of external data.[92] In February 2023, the Colorado Department of Regulatory Agencies Division of Insurance held a stakeholder meeting with a draft proposed Algorithm and Predictive Model Governance Regulation.[93] The details of how the regulation will be enforced are yet to be released. Rhode Island and Oklahoma had proposed bills patterned on Colorado's law, which both failed in committee.[94]

The current regulatory climate, in which state and local lawmakers are taking the initiative to promote algorithmic accountability, is a hopeful development. But it carries certain risks. While such efforts may help to fill the void left by federal inaction, the proliferation of decentralized regulation presents challenges federal regulation would not. A patchwork of inconsistent laws and regulations across states could be challenging for enforcement, industry, and consumers. Nonetheless, the growing momentum at the state level could be an important step toward ensuring that people are adequately protected. There is a pressing need for pragmatic approaches to implementing these laws, especially algorithmic audits, which will likely remain in the remit of local regulators.

2. Cases

As anticipated, the increasing prevalence of algorithms in decision-making has given rise to several new cases alleging discrimination,[95] further underscoring the need for effective regulation and accountability measures. Progress on such issues through the courts has been slow, although it is likely that pending cases will clarify some legal reasoning on algorithmic accountability this year.

In *Connecticut Fair Housing Center v. CoreLogic*, a pending case before a federal district court in Connecticut, algorithmic discrimination in housing is at issue.[96] The challenge relates to whether

---

consumer data and information sources, as well as algorithms and predictive models using external consumer data and information sources, which use has the result of unfairly discriminating based on race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression").

[91] *Id.* at §7(b)(I).

[92] *Id.*

[93] Colo. Dep't of Regul. Agencies Div. of Ins., Proposed Regulation on Algorithm and Predictive Model Governance (Feb. 1, 2023), https://doi.colorado.gov/announcements/for-review-and-comment-draft-proposed-algorithm-and-predictive-model-governance.

[94] R.I., H.B. 7230, Gen. Assemb., Jan. Sess. (R.I. 2022) (died in committee, recommended hold for further study); Okla., Insurance Consumer Rights Act H.B. 3186, 58th Leg., 2nd Sess., (2022) (died in committee.)

[95] *See, e.g.*, Mobley v. Workday, Inc., No. 23-cv-00770 (N.D. Cal. Feb. 21, 2023) a civil rights class action was filed against Workday, Inc. for alleged discrimination in its AI employment systems and screening tools.

[96] Conn. Fair Hous. Ctr. v. CoreLogic Rental Prop. Sols., No. 3:18-cv-705-VLB (D. Conn. Mar. 3, 2022).

an automated tenant screening service, if shown to be discriminatory, can be held responsible under the Fair Housing Act. At an earlier hearing on August 7, 2020, the District Court of Connecticut found there was sufficient statistical evidence to dispute whether CoreLogic's criminal background reporting system had a disparate impact and held that CoreLogic was unlikely to have a business justification for screening applicants solely on the basis of whether someone has a pending arrest.[97] The trial was heard in October and November 2022, and a ruling is pending. It will hopefully reveal more details about what is permitted or prohibited in algorithmic screening.

In January 2023, the Department of Justice (DOJ) and the Department of Housing and Urban Development (HUD) jointly filed a statement of interest in a similar pending lawsuit, *Louis et al. v. SafeRent Solutions LLC*, alleging discrimination against Black and Hispanic rental applicants based on the use of an algorithm-based tenant screening system.[98] The plaintiffs argue that the screening algorithm relies on factors that disproportionately disadvantage Black and Hispanic applicants, such as credit history and non-tenancy related debts, and fails to consider that the use of HUD-funded housing vouchers makes such tenants more likely to pay their rents.[99] In particular, the DOJ and HUD emphasize that housing providers and tenant screening companies that use algorithms to screen tenants are not exempt from FHA liability when their practices disproportionately deny people of color access to fair housing opportunities.[100]

There are more would-be 'test cases' that are waiting in the wings. For instance, Kyle Behm submitted a complaint to the Equal Employment Opportunity Commission (EEOC) alleging that pre-employment personality tests discriminate against applicants with certain mental disabilities.[101] Kyle, a young man with bipolar disorder, completed an online application process that, like many others, included the Kronos/Unicru personality tests, which 'red-flagged' Kyle.[102] However, those investigations with the EEOC are still ongoing.[103] Clarification of issues such as

---

[97] Conn. Fair Hous. Ctr. v. CoreLogic Rental Prop. Sols., 478 F. Supp. 3d 259 (D. Conn. 2020).
[98] Statement of Interest of the United States, Louis et al. v. SafeRent Solutions LLC, No. Case 1:22-cv-10800-AK (Mass. Dist. Ct. Jan. 9, 2023).
[99] Statement of Interest of the United States, Louis et al. v. SafeRent Solutions LLC, No. Case 1:22-cv-10800-AK (Mass. Dist. Ct. Jan. 9, 2023); Department of Justice, *Press Release, Justice Department Files Statement of Interest in Fair Housing Act Case Alleging Unlawful Algorithm-Based Tenant Screening Practices*, (Jan. 9, 2023), https://www.justice.gov/opa/pr/justice-department-files-statement-interest-fair-housing-act-case-alleging-unlawful-algorithm (last visited Apr. 10, 2023).
[100] *Id.*
[101] Although, one matter had a negotiated settlement resulting in a modification of online recruitment testing, see *Press Release, Bazelon Center for Mental Health Law, Lowe's Announces Changes to Online Application Process for Retail Employees* (Nov. 17, 2017), http://www.bazelon.org/wp-content/uploads/2017/11/Joint-Statement-with-Lowes.pdf.
[102] O'NEIL, *supra* note 2; Kelly Cahill Timmons, *Pre-Employment Personality Tests, Algorithmic Bias, and the Americans with Disabilities Act*, 125 PENN ST. L. REV. 389 (2021); Cathy O'Neil, *Personality Tests Are Failing American Workers*, BLOOMBERG (Jan. 18, 2018), https://www.bloomberg.com/opinion/articles/2018-01-18/personality-tests-are-failing-american-workers; Lauren Weber & Elizabeth Dwoskin, *Are Workplace Personality Tests Fair?*, WALL STREET JOURNAL (Sept. 30, 2014), http://online.wsj.com/articles/are-workplace-personality-tests-fair-1412044257.
[103] Timmons, *supra* note 102 (citing an interview with Kyle's father, Roland Behm).

those raised by this case, including liability for businesses using third-party algorithmic tools and the potential disparate impact of automated tools, would be a meaningful step towards understanding how anti-discrimination law applies to algorithmic systems.

3. Regulatory action

Federal regulators are ramping up initiatives in relation to automated decision-making. In 2022, Meta settled with the DOJ over claims of discrimination in its housing advertisements, providing an important example of a regulatory action that addresses algorithmic fairness head on.

As part of an investigation, the news organization ProPublica ran housing Facebook advertisements in 2016 that excluded non-white people from seeing the ad.[104] After more investigations by news organizations, the DOJ began its own investigation that led to a settlement in 2022.[105] In the settlement, Meta agreed to keep tabs on housing ads with a so-called "variance reduction system" that would make sure the demographic that eventually sees an ad campaign does not differ substantially from the "eligible audience" for that ad. The eligible audience, which is defined by the advertiser, has also been restricted so that it cannot exclude people by race or other protected class categories.

There are a few characteristics of the Meta-DOJ settlement that are important and, as will be expanded upon, are shared with the Explainable Fairness framework outlined in this Article. First, the settlement defines a metric that addresses a specific harm, namely unequal targeting and delivery of certain advertisements.[106] Second, it defines what level of difference is considered significant–in other words, it defines a threshold that Meta is meant to stay under in order to be compliant with the settlement's anti-discrimination provisions.[107] Third, it requires the continuous monitoring of this fairness metric, as well as regular reporting on adherence to that metric to the DOJ.

---

[104] Julia Angwin & Terry Parris Jr, *Facebook Lets Advertisers Exclude Users by Race*, PROPUBLICA (Oct. 28, 2016), https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race.

[105] Settlement Agreement and Final Judgment, United States v. Meta Platforms, Inc., No. 1:22-cv-05187, 22 Civ. 5187 (S.D.N.Y. June 27, 2022); Department of Justice, Press Release, United States Attorney Resolves Groundbreaking Suit Against Meta Platforms, Inc., Formerly Known As Facebook, To Address Discriminatory Advertising For Housing, https://www.justice.gov/usao-sdny/pr/united-states-attorney-resolves-groundbreaking-suit-against-meta-platforms-inc-formerly (last visited Apr. 11, 2023).

[106] Complaint & Demand for Jury Trial, United States v. Meta Platforms, Inc., No. 1:22-cv-05187 (S.D.N.Y. filed June 21, 2022) (explaining that the DOJ had conducted extensive testing, which controlled for sex, age, and income, that showed statistically significant racial disparities in Facebook's ad delivery).

[107] Joint Letter addressed to Judge John G. Koeltl, United States v. Meta Platforms, Inc., 22 Civ. 5187 (JGK) (S.D.N.Y., Jan. 9, 2023) (informing the Court that the parties had reached final agreement on the Variance Reduction System to reduce variances in the delivery of housing ads that the DOJ alleges are introduced by Meta's ad delivery system, for sex and estimated race/ethnicity pursuant to the Settlement Agreement)

Another advantageous provision of this settlement is that it places the burden on Meta to do the technical work of making things fair.[108] As a result, lawyers at the DOJ do not have to understand the internal engineering of what is required to make housing ads non-discriminatory, just evaluate the evidence Meta presents to conclude if Meta is complying with the terms of the settlement.

Similarly, HUD tried to implement a new rule for algorithmic disparate impact that was recently overturned.[109] Despite intentions, in practice the rule would have allowed a defense to algorithmic disparate impact that was unjustified. The proposed 2020 Disparate Impact Final Rule attempted to amend HUD's interpretation of the FHA's disparate impact standard in order to better reflect *Inclusive Communities.* In addition to introducing five elements a plaintiff must establish in a disparate impact claim, the rule proposed three defenses for defendants who are accused of causing disparate impact in their use of algorithmic models. The first defense excused a defendant who "[p]rovides the material factors that make up the inputs used in [a] challenged model and shows that these factors do not rely in any material part on factors that are substitutes or close proxies for protected classes under the Fair Housing Act and that the model is predictive of credit risk or other similar valid objective."[110] The second defense indemnifies a defendant against disparate impact liability if the defendant "[s]hows that the challenged model is produced, maintained, or distributed by a recognized third party" whose tool is standard in the industry "and [that] the defendant is using the model as intended by the third party."[111] Finally, the third defense excuses a defendant that "shows that the model has been subjected to critical review and has been validated by . . . [a] third party that has analyzed the challenged model and found that the model . . . accurately predicts risk or other valid objectives, and that none of the factors used in the algorithm rely . . . on factors that are substitutes or close proxies for protected classes under the Fair Housing Act."[112]

Courtesy of a preliminary injunction in the U.S. District Court for the District of Massachusetts,[113] and a change in administration,[114] the widely criticized rule was replaced in 2023.[115] In replacing the prior rule, HUD responded to the previous algorithmic defense saying it "could in practice improperly exempt many housing-related practices that are increasingly reliant upon algorithms

---

[108] Settlement Agreement and Final Judgment, United States of America v. Meta Platforms, Inc., No. 1:22-cv-05187, 22 Civ. 5187 (S.D.N.Y., June 27, 2022).
[109] *See* HUD's Implementation of the Fair Housing Act's Discriminatory Impact Standard, 85 Fed. Reg. 60288 (Sept. 24, 2020) (to be codified at 24 C.F.R. pt. 100).
[110] HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42854, 42859 (proposed Aug. 19, 2019).
[111] *Id.* at 42862.
[112] *Id.* at 42862.
[113] Mass. Fair Hous. Ctr. v. U.S. Dep't of Hous. & Urban Dev., 496 F. Supp. 3d 600 (D. Mass. 2020).
[114] The White House, *Memorandum on Redressing Our Nation's and the Federal Government's History of Discriminatory Housing Practices and Policies*, THE WHITE HOUSE, https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/26/memorandum-on-redressing-our-nations-and-the-federal-governments-history-of-discriminatory-housing-practices-and-policies/ (last visited Apr. 10, 2023).
[115] Reinstatement of HUD's Discriminatory Effects Standard, 88 Fed. Reg. 19450 (to be codified at 24 C.F.R. pt. 100) (Department of Housing and Urban Development Mar. 31, 2023).

and automated processes that rely on outcome predictions, such as lending practices, from liability under a disparate impact standard."[116] The return to the 2013 rule will likely alleviate various concerns about the proposed 2020 rule.[117]

Other federal agencies are allocating resources to pursue similar actions.[118] State insurance regulators have taken to issuing notices and circular letters highlighting concerns of algorithmic bias and discrimination in insurance.[119] Well-resourced regulators, with strong authority as well as capacity, will be an essential part of the enforcement and compliance of algorithms with protections for individuals affected by algorithms.

4. Voluntary industry practices

Without clear guidance on how to comply with the wide variety of emerging regulatory approaches businesses are improvising ways to address algorithmic discrimination. One way they have approached the issue is through voluntary "exhibition" audits. In 2018, Facebook (now Meta)

---

[116] *Id.* at 19493; *see also* Department of Housing and Urban Development, *Press Release, HUD Restores "Discriminatory Effects" Rule*, HUD, https://www.hud.gov/press/press_releases_media_advisories/hud_no_23_054 (last visited Apr. 10, 2023).

[117] Foggo and Villasenor, *supra* note 51; John Villasenor and Virginia Foggo, *Why a Proposed HUD Rule Could Worsen Algorithm-Driven Housing Discrimination*, BROOKINGS (Apr. 16, 2020), https://www.brookings.edu/blog/techtank/2020/04/16/why-a-proposed-hud-rule-could-worsen-algorithm-driven-housing-discrimination/; Tracy Jan, *HUD Raises the Bar for Bringing Discrimination Claims*, WASHINGTON POST (Aug. 17, 2019), https://www.washingtonpost.com/business/2019/08/16/hud-raises-bar-bringing-discrimination-claims/.

[118] Stephanie Nguyen, *A Century of Technological Evolution at the Federal Trade Commission*, FEDERAL TRADE COMMISSION (2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/02/century-technological-evolution-federal-trade-commission (last visited Feb 20, 2023) (establishing the FTC Office of Technology); Consumer Financial Protection Bureau, *CFPB Launches New Effort to Promote Competition and Innovation in Consumer Finance*, (2022), https://www.consumerfinance.gov/about-us/newsroom/cfpb-lauches-new-effort-to-promote-competition-and-innovation-in-consumer-finance/ (last visited Feb 24, 2023) (establishing the CFPB Office of Competition and innovation); Equal Employment Opportunity Commission, *EEOC Launches Initiative on Artificial Intelligence and Algorithmic Fairness*, (2021), https://www.eeoc.gov/newsroom/eeoc-launches-initiative-artificial-intelligence-and-algorithmic-fairness (last visited Feb 20, 2023).

[119] Conn. Ins. Dep't, Notice Concerning the Usage of Big Data and Avoidance of Discriminatory Practices (Apr. 20, 2022), https://portal.ct.gov/-/media/CID/1_Notices/Technologie-and-Big-Data-Use-Notice.pdf (requiring insurers to comply with applicable anti-discrimination laws and complete annual data certification by Sept. 1, 2022); Cal. Ins. Comm'r, Bulletin 2022-5 Concerning Allegations of Racial Bias and Unfair Discrimination in Marketing, Rating, Underwriting, and Claims Practices by the Insurance Industry (June 30, 2022), https://www.insurance.ca.gov/0250-insurers/0300-insurers/0200-bulletins/bulletin-notices-commiss-opinion/upload/BULLETIN-2022-5-Allegations-of-Racial-Bias-and-Unfair-Discrimination-in-Marketing-Rating-Underwriting-and-Claims-Practices-by-the-Insurance-Industry.pdf (highlighting civil rights and insurance laws protecting against discrimination and unfair treatment, and noting concern over allegations of insurers flagging claims based on zip code or biometric data); N.Y. Dep't of Fin. Servs., Ins. Circular Letter No. 1 Concerning Use of External Consumer Data and Information Sources in Underwriting for Life Insurance (Jan. 18, 2019), https://www.dfs.ny.gov/industry_guidance/circular_letters/cl2019_01 (explaining insurers should not use external data, algorithms or predictive models unless they have determined such tools do not use prohibited criteria, are not unfairly discriminatory, and if statistical data is used, there is a valid rationale).

commissioned a civil rights audit, resulting in a public report.[120] In 2020, Twitter conducted a review of race-bias in Twitter's algorithm to auto-crop images.[121] That year, Airbnb launched Project Lighthouse to collect data needed to measure and evaluate discrimination on its platform in the US.[122]

Other companies have participated in industry-level development of policy or standards with respect to algorithmic bias.[123] There has also been an emergence of other self-regulatory efforts such as voluntary AI ethics frameworks, many of which are produced by private companies.[124] However, AI ethics frameworks cannot address a company acting without accountability, leaving affected individuals with no way of challenging an AI-informed decision.[125] IBM's AI Fairness 360,[126] Amazon Sagemaker,[127] and other tools purport to visualize or even mitigate bias but offer no suggestions on what metric exposes unfair discrimination, how much difference in outcomes for different groups is too much, or what factors might explain or legitimize those differences.[128] A wide range of experts have rejected self-regulation as a sufficient tool to protect against discrimination and other human rights infringements in the context of algorithmic decision-making.[129] Industry self-regulation to establish voluntary protections against bias needs to be

---

[120] Laura Murphy, Facebook's Civil Rights Audit - Final Report, Relman Colfax (July 8, 2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf.

[121] Kyra Yee, Uthaipon Tantipongpipat & Shubhanshu Mishra, *Image Cropping on Twitter: Fairness Metrics, Their Limitations, and the Importance of Representation, Design, and Agency*, 5 PROC. ACM HUM. COMPUT. INTERACT. 1 (2021), http://arxiv.org/abs/2105.08667; Rumman Chowdhury, *Sharing Learnings about Our Image Cropping Algorithm*, TWITTER INSIGHTS (May 19, 2021), https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.

[122] SID BASU ET AL., *Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data*, (2022), https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf; Airbnb, *Measuring Discrimination on the Airbnb Platform*, AIRBNB NEWSROOM (June 15, 2020), https://news.airbnb.com/measuring-discrimination-on-the-airbnb-platform/.

[123] Rachel DuFault, Algorithmic Bias Is No Longer Under Regulators' Radar, BLOOMBERG LAW (Nov. 14, 2022), https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-algorithmic-bias-is-no-longer-under-regulators-radar; Center for Industry Self-Regulation, *AI in Hiring & Recruiting*, BBB NATIONAL PROGRAMS CHARITABLE FOUNDATION, https://industryselfregulation.org/incubator/ai-hiring (last visited Apr. 10, 2023); *Algorithmic Bias Safeguards for Workforce*, DATA & TRUST ALLIANCE, https://dataandtrustalliance.org/ (last visited Apr. 10, 2023).

[124] Anna Jobin, Marcello Ienca & Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, 1 NAT MACH INTELL 389 (2019) (demonstrating 22.6% of AI ethics frameworks identified were published by private companies); Ray Eitel-Porter, *Beyond the Promise: Implementing Ethical AI*, 1 AI ETHICS 73 (2021); Ben Wagner, *Ethics as an Escape from Regulation: From "Ethics-Washing" to Ethics-Shopping?*, *in* BEING PROFILED: COGITAS ERGO SUM (Emre Bayamlioglu et al. eds., 2018).

[125] AUSTRALIAN HUMAN RIGHTS COMMISSION, HUMAN RIGHTS AND TECHNOLOGY 90 (2021).

[126] *AI Fairness 360*, IBM RESEARCH, https://aif360.mybluemix.net/aif360.mybluemix.net (last visited Apr. 11, 2023).

[127] *Sagemaker*, AMAZON WEB SERVICES, https://aws.amazon.com/sagemaker/ (last visited Apr. 11, 2023).

[128] Michelle Seng Ah Lee & Jatinder Singh, *The Landscape and Gaps in Open Source Fairness Toolkits*, *in* CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (ACM 2021) https://dl.acm.org/doi/pdf/10.1145/3411764.3445261.

[129] Lee and Singh, *supra* note 129; Karen Yeung, Andrew Howes & Ganna Pogrebna, *AI Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing*, *in* THE OXFORD HANDBOOK OF

complemented with concrete federal and state mandates about algorithmic bias with robust enforcement regimes and budgets.

D. Gaps in current work

There remain considerable challenges in understanding and implementing statistical metrics for fairness proposed in algorithmic fairness literature. Such work is helpful in guiding technical ways to identify disparity in outcomes for different individuals subject to algorithms, although further work is needed to align this contribution in the legal compliance of algorithms.

Anti-discrimination laws may be difficult to implement in complex algorithmic contexts. As discussed above, applying disparate impact doctrine to algorithmic contexts will be difficult for three primary reasons: (1) it is unclear what the threshold will be to discern significant disparity for a prima facie case; (2) it may be difficult to identify what variables reflect a legitimate business interest; (3) it is futile for plaintiffs to identify less discriminatory alternatives in the context of an optimized algorithm. Further, prominent areas of algorithmic use may be unregulated by disparate impact doctrine. Moreover, even when the doctrine does apply, the burden it places on plaintiffs is often prohibitively high, leading to a very low success rate for plaintiffs.[130] That success rate would not improve with those who seek to challenge discriminatory algorithms.

In response to these gaps, this Article proposes a new approach, one that takes inspiration from existing law but also charts new territory. The proposed Explainable Fairness framework aims to provide clearer guidelines for understanding, measuring, and remedying disparities in algorithmic decisions, with potential applications extending beyond current disparate impact doctrine.

III. Our Proposal

A. Explainable Fairness framework

In response to the challenges identified in respect of algorithmic fairness and discrimination, this Article proposes the Explainable Fairness framework.

The plaintiff's burden in establishing a disparate impact claim, especially their obligation to find an equally viable yet less discriminatory alternative, will be an insurmountable hurdle to hold discriminatory algorithms to account. Further, there is a larger hurdle that disparate impact is not

---

ETHICS OF AI 75, 77 (Markus D. Dubber, Frank Pasquale, & Sunit Das eds., 2020); Evgeni Aizenberg & Jeroen van den Hoven, *Designing for Human Rights in AI*, 7 BIG DATA & SOCIETY 1 (2020).

[130] Selmi, *supra* note 49, at 738–39 (demonstrating that on average plaintiffs had a 19.2% success rate in the reported appellate cases for 1984-85, 1994-95, and 1999-2001, and a 25.1% success rate in the in the district court decisions for six years (1983, 1987, 1991, 1996, 1999, 2002)).

actionable in some contexts, including the insurance setting which is a prominent space for algorithmic decision-making. If a disparate impact claim is unlikely to succeed, a new approach should be taken to overcome the material risk to individuals subject to algorithms across all areas of life.

The Explainable Fairness framework is aimed at regulators to implement a discrimination analysis. It offers a structured way to identify harms and related outcomes of interest, define significant disparities in those outcomes, and to assess whether disparities can be explained by legitimate (non-discriminatory) reasons. Importantly, the framework decouples technical questions of data science from regulatory and legal questions.

In proposing something new, regulators can be empowered to flip the burden on discrimination and require regulated entities to provide explanations and engage in negotiations if it appears some consumers are being penalized or treated differently by algorithmic systems. The remainder of this Article demonstrates how this framework can draw from disparate impact analysis and other relevant laws but be something quite distinct to take a new approach to new challenges.

1. *How Explainable Fairness broadly applies*

The Explainable Fairness framework addresses the context of a regulator defining compliance of algorithmic systems with respect to a given law. The goal is to define and measure fair treatment, under a specific law, for a specific protected class and with respect to a specific algorithmic system. Therefore, an employment regulator could conduct a fair hiring analysis, or an insurance regulator could conduct a review of unfair race or sex based discrimination.

If a regulator suspects unlawful discrimination arising from an algorithmic system, the Explainable Fairness framework proceeds as follows. First, the regulator identifies the relevant stakeholder groups, which will include specific protected classes as defined by the discrimination laws in question and specifies certain outcomes of interest that are within the scope of the law.[131]

A standard scenario is that the algorithmic system issues a risk score for each impacted person, which is taken into consideration when making a consequential decision about that person, such as whether they are offered insurance, or asked to interview for a job. Then, the regulator may be concerned if members of a legally protected class (i.e., a stakeholder group) are being issued inflated scores (i.e., a disparity in an outcome of interest). Assuming that membership to protected class is either known or reasonably inferable, a straightforward analysis of raw average scores per protected class can give a high-level description of how members of each class are being scored.

[131] There are many ways the regulator could identify stakeholder groups and outcomes of interest; the Ethical Matrix framework is one option, *see infra* Section III.B.

If substantial differences in average scores are discovered, then the regulator and companies using these scoring systems engage in a structured, iterative process of interrogating and explaining the differences. The company can argue that the differences are due to differences across protected classes in some relevant characteristics of affected persons. For instance, a company using a hiring algorithm could argue that women got better scores than men because they had more years of experience on average and thus were better qualified for the job. The regulator's role includes defining the process and rules for the negotiation that establishes whether a given factor is legitimate in that algorithmic context, and if so, exactly how it should be accounted for in measuring differences between groups in outcomes.

In this way, fairness is determined not as a formula, but rather as part of a negotiated and contextual process, adherent to the law and to industry practices. The ultimate question the Explainable Fairness framework aims to address is whether the system treats individuals of varying protected classes similarly. Here, similarity is measured quantitatively based on agreed metrics and legitimate factors associated with the identified outcomes. This analysis yields a comprehensible, plain-English result, clarifying that individuals with similar attributes and circumstances are being treated comparably. Thus, the Explainable Fairness framework aids in understanding and evaluating the fairness of algorithmic systems.

2. *How Explainable Fairness identifies statistical significance*

Drawing from disparate impact's requirement for significant differences in establishing a prima facie case, this Article proposes a new way to define substantial differences in the Explainable Fairness framework.

The regulator initially employs a naive definition, considering any non-zero difference in outcomes significant enough for an inquiry. They can collect data from entities within their jurisdiction, forming a snapshot of industry performance regarding outcome differences. This "consumer reports view" allows the regulator to iteratively define what constitutes a substantial difference, relative to the industry.[132] This definition evolves in response to industry shifts, creating a positive feedback loop that stimulates continual improvement in fairness standards.

An iteratively updated definition encourages constant improvement within the industry, establishing a positive feedback loop where compliance drives up the median standards and incentivizes lagging companies to adapt. This not only progressively raises the bar for everyone but also rewards those companies that prioritize fairness.

[132] An industry snapshot would show all regulated entities ranked according to outcome differences, such as the Consumer Reports reviews, see, What we do, CONSUMER REPORTS (2023), https://www.consumerreports.org/cro/about-us/what-we-do/index.htm (last visited Jul 10, 2023).

This process of defining substantial is inherently a policy and moral issue, not solely a technical one. It's important to recognize that algorithm-generated disparities often mirror historical and ongoing social inequities reflected in data used to train these algorithms.[133] Past and current injustice "very plausibly caused the observed differences between and among protected groups".[134] There is a material risk that algorithms will amplify historical injustice because of the information reflected in biased datasets.[135] Algorithms may then result in "compounding injustice" by harming an individual or group that has already been victimized.[136] Therefore, regulators must carefully design their iterative definition process to avoid being informed by past and ongoing injustices.

In addition to the moral and policy considerations of disparity, legal precedent may in some cases inform a regulator's development of the definition. Courts have not drawn clear lines on what disparities are sufficiently significant to establish legal liability.[137] They are concerned with a disparity that is statistically significant – meaning the disparity is less likely due to chance. While some disparities will be evidently significant without the need for statistical evidence, others conversely will be so small they can be comfortably rejected. However, the significance of disparities may be difficult to define in some cases, so guidance can be drawn from both the courts and federal agency guidelines. For example, in the context of employment discrimination, longstanding EEOC guidance provides a "rule of thumb" that the impact ratio (i.e., the selection rate among this group, divided by the selection rate among the most-selected group in this protected class) should be at least four-fifths.[138] However, this rule is not dispositive.[139]

Legal precedents and federal agency guidelines may also guide regulators' development of definitions. While courts haven't delineated clear boundaries for disparities significant enough to establish legal liability, they have shown concern for statistically significant disparities. Longstanding EEOC guidance in employment discrimination, for example, provides a "rule of thumb" for impact ratios.[140] Courts have also considered standard deviations in expected versus

---

[133] AUSTRALIAN HUMAN RIGHTS COMMISSION, *Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias*, (2020), https://humanrights.gov.au/our-work/rights-and-freedoms/publications/using-artificial-intelligence-make-decisions-addressing.

[134] Deborah Hellman, *The Epistemic Commitments of Nondiscrimination*, 4 *in* OXFORD STUDIES IN PHILOSOPHY OF LAW 156, 164 (John Gardner, Leslie Green, & Brian Leiter eds., 2021).

[135] O'NEIL, *supra* note 2.

[136] Deborah Hellman, *Indirect Discrimination and the Duty to Avoid Compounding Injustice*, *in* FOUNDATIONS OF INDIRECT DISCRIMINATION LAW 105, 114 (Hugh Collins & Tarunabh Khaitan eds., 2018).

[137] Clady v. Cty. of Los Angeles, 770 F.2d. 1421 (1985); citing Smith v. Xerox Corp., 196 F.3d. 358 (1999) ("[T]he substantiality of a disparity is judged on a case-by-case basis."); Groves v. Alabama State Bd. of Educ., 776 F. Supp. 1518 (1991) ("There is no rigid mathematical threshold that must be met to demonstrate a sufficiently adverse impact.").

[138] Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11996 (Mar. 2, 1979) (to be codified at 29 CFR pt. 1607).

[139] 28 C.F.R. § 50.14(4)(D).

[140] Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11996 (Mar. 2, 1979) (to be codified at 29 CFR pt. 1607); U.S. Equal Employment Opportunity Commission, *supra* note 70.

observed rates.[141] More recently, social science standards of statistical significance, such as the Shoben formula, have been applied.[142] Nevertheless, these rules and formulas are not definitive, and regulators must navigate this complex issue with consideration for the policy, moral, and legal dimensions.

No particular method or threshold is the default rule. Therefore, defining a statistically significant or big disparity can be aided by this relative and iterative process.

3. *How Explainable Fairness engages with legitimate factors*

In general, the law takes a formally neutral view that expects no less favorable treatment based on protected attributes. Indeed, anti-discrimination doctrine generally prohibits the intentional use of variables that are protected characteristics.[143] However, there are some exceptions where there is a connection to true risk or inherent differences between individuals.[144] There is, for instance, a legitimate distinction to make with respect to the inherent differences between men and women in respect of reproductive issues.[145] In another context, most state insurance rules in the US allow that a driver with many prior accidents is charged more for car insurance than a driver with no prior accidents, all else being equal.[146] One of the regulator's key tasks in the Explainable Fairness framework is to establish a set of "legitimate factors," or a process for assessing whether a given factor proposed by a regulated entity as an explanation for disparate outcomes is "legitimate."

Currently, there are ways through legal precedent, regulatory frameworks, and principles for regulators to ascertain which factors may be contextually legitimate. Consider the Equal Credit Opportunity Act that protects against discrimination because of an applicant's income from any public assistance program, in addition to discrimination based on protected class.[147] This is because the "receipt of public assistance benefits [and some other factors]…however, Congress deemed it legitimate to take those factors into account under certain circumstances."[148] Also, under the

---

[141] Hazelwood School District v. United States, 433 U.S. 299 (1977).
[142] Groves v. Alabama State Bd. of Educ., 776 F. Supp. 1518, 1526–28 (1991); citing Richardson v. Lamar Cty. Bd. of Educ., 729 F. Supp. 806, 816 (1989) (the "Shoben formula" recognizes a "Z-value" measuring the difference in the groups' success rates greater than 1.96 standard deviations to be statistically significant).
[143] *See, e.g.,* prohibitions in the employment context, 42 U.S.C. § 2000e; lending, 15 USC § 1691(a)(1)–(3); and housing, 42 USC § 3605(a).
[144] Holli Sargeant & Måns Magnusson, *Automated Decisions and Anti-Discrimination Doctrine* (2023) (unpublished manuscript) (on file with Author).
[145] United States v. Virginia 518 US 515 (1996); see discussion in Deborah Hellman, *Sex, Causation, and Algorithms: How Equal Protection Prohibits Compounding Prior Injustice*, 98 WASH. U. L. REV. 481 (2020).
[146] *See* National Association of Insurance Commissioners, *Auto Insurance*, NAIC (2023), https://content.naic.org/cipr-topics/auto-insurance; Paul E. Patterson & Matthew Kuofie, *Statistical Analysis of Crash Factors Related to Auto Insurance*, 11 J. FIN., ACCT. & MGMT. 25 (2020); Xi Xin & Fei Huang, *Antidiscrimination Insurance Pricing: Regulations, Fairness Criteria, and Models*, N. AM. ACTUARIAL J. 1 (2023).
[147] 15 U.S.C. §1691(a)(2).
[148] PATRICIA A. MCCOY, BANKING LAW MANUAL: FEDERAL REGULATION OF FINANCIAL HOLDING COMPANIES, BANKS AND THRIFTS § 8.02[1][a][ii] (2 ed. 2015).

Affordable Care Act, health insurance companies were allowed to vary premiums based on age and smoking status, but within specific limits (at most 3:1 based on age, and at most 1.5:1 based on smoking status).[149] In cases that the law considers, the regulator's task is simpler. However, things get more complicated when a proposed explanatory factor is correlated with protected class status. In the context of predictive models such factors are sometimes called "proxy variables."[150] When a proxy variable is included in a predictive model, it can cause disparities across protected classes in the model's predictions (e.g., risk scores) precisely because it contains information about each person's protected class status. Such practice is often called *proxy discrimination*.[151] Schwarcz and Prince explain that "a practice producing a disparate impact only amounts to proxy discrimination when the usefulness to the discriminator of the facially neutral practice derives, at least in part, from the very fact that it produces a disparate impact."[152] They explain that in algorithmic contexts, "proxy discrimination need not be intentional when membership in a protected class is predictive of a discriminator's facially neutral goal, making discrimination 'rational.'"[153]

A facially neutral variable may result in disparate impact to a protected group but be legally permissible to consider (under disparate impact doctrine) if it is connected with a legitimate business interest and there is no less discriminatory alternative. Whether or not it is legally permissible to consider a given variable in making a prediction, the discriminatory effects that may be produced by such algorithmic practices could amount to unfairness that the regulator believes warrants further inquiry.

This Article contributes to existing literature methods in which a regulator may make structured decisions about what constitutes a legitimate factor in specific regulated contexts.

Balancing test

---

[149] Patient Protection and Affordable Care Act, Pub. L. No. 111-148, 124 Stat. 119 (2010) (codified as amended at 42 U.S.C. § 18001); Market Rules and Rate Review Final Rule, 45 C.F.R pt. 147 (2013); U.S. Centers for Medicare & Medicaid Services, *Market Rating Reforms*, https://www.cms.gov/CCIIO/Programs-and-Initiatives/Health-Insurance-Market-Reforms/Market-Rating-Reforms (last visited Apr. 18, 2023).

[150] There is still a risk in algorithmic design that humans may model an algorithm to use a facially neutral classification to serve as an intentional proxy for a protected characteristic, *see* Hellman, *supra* note 19, at 851.

[151] See Gillis and Spiess, *supra* note 19, at 470; Barocas and Selbst, *supra* note 24, at 695; Deborah Hellman, *Defining Disparate Treatment: A Research Agenda for Our Times*, 99 IND. L.J. forthcoming (2023); Raphaële Xenidis, *Tuning EU equality law to algorithmic discrimination: Three pathways to resilience*, 27 MAASTRICHT J. EUR. & COMP. L. 736, 745 (2020).

[152] Daniel Schwarcz & Anya Prince, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1260 (2020).

[153] *Id.* at 1262.

A quantitative balancing test is one potential method to evaluate the legitimacy of a factor.[154] This involves comparing the strength of the factor's prediction of a crucial business outcome with its correlation with protected class status. In general, a factor's legitimacy increases with its predictability and decreases with its association with protected class status. A ratio test exemplifies this approach by validating factors that are stronger predictors than they are proxies for protected class status.[155]

However, this balancing test does not account for the underlying reasons for a factor's correlation with business outcomes or protected class status, which may influence fairness perceptions. For instance, a factor may have limited predictive power but a strong link to race, leading to diverging views on its legitimacy depending on whether its predictive basis is considered related or unrelated to race.

Causal story

The ratio test sketched in the Appendix, is one example of such a quantitative balancing test. It would permit legitimate factors that are relatively strong as predictors, compared to their strength as proxies for protected class status.[156]

The legitimacy of a factor is highly dependent on its specific context and can differ based on varying scenarios. To assess the legitimacy of a factor, a regulator might examine whether there is a "causal story" linking the model's objective and the utilized feature.[157]

Take life insurance as an example, which heavily relies “on the life expectancies of individuals at the insured's age”.[158] Because life insurance triggers upon the policyholder’s death, the heavy

---

[154] It should be noted that the term “balancing test” used here does not refer to legal balancing tests in the traditional sense, i.e., it does not refer to the process by which courts weigh various factors or considerations in order to determine the outcome of certain types of legal disputes, see, e.g., Patrick M. McFadden, *The Balancing Test*, 29 B.C. L. REV. 585 (1987).

[155] See similar cost-benefit ratios in Richard Zerbe & Scott, Tyler, *A Primer for Understanding Benefit-Cost Analysis*, *in* ACTIONABLE INTELLIGENCE FOR SOCIAL POLICY: USING INTEGRATED DATA SYSTEMS TO ACHIEVE A MORE EFFECTIVE, EFFICIENT, AND ETHICAL GOVERNMENT (John Fantuzzo, Dennis Culhane, & Heather Rouse eds., 2015); Radhika Bhula, Meghan Mahoney, & Kyle Murphy, *Conducting Cost-Effectiveness Analysis (CEA)*, THE ABDUL LATIF JAMEEL POVERTY ACTION LAB (J-PAL) (2020), https://www.povertyactionlab.org/resource/conducting-cost-effectiveness-analysis-cea. *See* discussions *infra* Appendix; *see* discussions *infra* Appendix.

[156] It also imposes an overall floor as to the predictiveness of a factor (i.e., non-predictive factors can’t be legitimate, no matter how little they proxy for protected class) and a ceiling on the extent to which legitimate factors jointly proxy for protected class (i.e., no set of factors can be legitimate if together they form a very strong proxy for protected class).

[157] Sargeant and Magnusson, supra note 145; Boyarskaya, Margarita et al., *What Is a Proxy and Why Is It a Problem? in* CONFERENCE ON FAIRNESS, ACCOUNTABILITY AND TRANSPARENCY (2022), https://www.youtube.com/watch?v=Qb0Q0HWBo1I.

[158] Kenneth S. Abraham, *Efficiency and Fairness in Insurance Risk Classification*, 71 VA. L. REV. 403, 418 (1985).

reliance on age as a classifying factor in life insurance is reasonable in this context. This suggests there could be sex-based differences in life insurance costs in the US, since life expectancy varies between men and women.[159] To explain race differences in life insurance costs, a regulator might consider factors such as smoking status. Black Americans have a higher likelihood of smoking, which increases mortality rates,[160] thus forming a "causal story" with the business-critical outcome of mortality, making it a more legitimate factor in this context.

Yet even the simplest causality approach faces significant fairness questions concerning agency. Individuals are embedded in complex causal chains that often extend beyond their control. For example, if tobacco products were more heavily advertised and sold in minority-dominated neighborhoods, the fairness of using smoking status to justify racial differences in life insurance pricing would be debatable.[161]

Precisely defining fairness in a specific, high-stakes context is complex, particularly given various normative conceptions of fairness.[162] This Article aims to demonstrate the regulator's role in defining fairness in its respective enforcement and the Explainable Fairness framework can be used to ensure regulator's decisions appropriately operationalize its legal and ethical choices.

B. Ethical Matrix: Measuring Harms

An important aspect of the Explainable Fairness framework is how to establish and measure potential harms. It is clear that algorithmic context has to be well-defined to even attempt to understand when it fails. If a scoring system accurately gauges one's risk of getting Type II diabetes in the next two years, it might be a useful tool in the hands of one's doctor, but a high-risk tool in the hands of companies where one might try to apply for a job. Context is everything.

An Ethical Matrix can be developed to define all stakeholders of an algorithmic system, consider their perspectives, and search for scenarios where the system fails or harms the stakeholder.[163]

---

[159] Although, the minority of females with a higher than average life expectancy "subsidizes" the minority of males who have lower than average life expectancies, see *Id.* at 435–6.

[160] *See* Jihyoun Jeon et al., Mortality Relative Risks by Smoking, Race/Ethnicity, and Education, 64 American Journal of Preventive Medicine S53 (2023), https://www.sciencedirect.com/science/article/pii/S0749379722005712; KFF, Adults Who Report Smoking by Race/Ethnicity, (2022), https://www.kff.org/other/state-indicator/smoking-adults-by-raceethnicity/ (last visited Jun. 22, 2023).

[161] *See* Jeon et al., supra note 161; KFF, supra note 161.

[162] Deirdre Mulligan et al., *This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology*, 3 *in* PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION 1 (2019), https://dl.acm.org/doi/10.1145/3359221; STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL (2019).

[163] Cathy O'Neil & Hanna Gunn, *Near-Term Artificial Intelligence and the Ethical Matrix*, *in* ETHICS OF ARTIFICIAL INTELLIGENCE 237 (S. Matthew Liao ed., 2020), https://doi.org/10.1093/oso/9780190905033.003.0009; See also, Céline Kermisch & Christophe Depaus, *The Strength of Ethical Matrixes as a Tool for Normative Analysis Related to Technological Choices: The Case of Geological Disposal for Radioactive Waste*, 24 SCI. ENG. ETHICS 29 (2018), https://doi.org/10.1007/s11948-017-9882-6; Ben Mepham, *Ethical Principles and the Ethical Matrix*, *in* PRACTICAL

Rows on the matrix represent different stakeholders, while columns reflect potential failures or successes from their perspective, gathered through interviews facilitated by a representative.[164] Interviews are conducted by a facilitator who explains the algorithmic system to appropriate representatives of stakeholder groups.[165] Some stakeholders, like future generations or the environment, cannot participate directly, necessitating inclusion of domain experts or advocates to identify potential risks. While it's challenging to capture every potential failure, the matrix is designed to highlight major existential threats and ethical trade-offs, such as contrasting one stakeholder group's desire to minimize false positives whereas another stakeholder group wants to minimize false negatives.[166]

The next step in this process is to grade risk within the matrix by assessing the probability and potential impact of each identified concern. This incorporates both likelihood and severity of a negative outcome and might prioritize differing harms based on the ethical stance of the stakeholders.[167] Some stakeholders may prioritize an approach that prioritizes the maximum utility for the greatest number of people.[168] Alternatively, if one adopts a deontological method, a review of harms would allow a discussion of whether the results of an algorithm are just.[169]

Risks are then ranked, and measures are established to monitor the occurrence of these harms, adjusting the algorithmic system as necessary to mitigate damage. The risk should be ranked from high to low and procedures should be established to measure the extent to which those harms are actually happening. Importantly, the Ethical Matrix approach is predominantly non-technical; the primary technical requirements come into play during risk monitoring and necessary algorithmic adjustments. This separates the task of defining ethical priorities and fairness from the technical

---

ETHICS FOR FOOD PROFESSIONALS 39 (2013), https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118506394.ch3; Karsten Klint Jensen et al., *Facilitating Ethical Reflection Among Scientists Using the Ethical Matrix*, 17 SCI. ENG. ETHICS 425 (2011), https://doi.org/10.1007/s11948-010-9218-2; Doris Schroeder & Clare Palmer, *Technology Assessment and the "Ethical Matrix,"* 1 POIESIS & PRAXIS: INTERNATIONAL JOURNAL OF TECHNOLOGY ASSESSMENT AND ETHICS OF SCIENCE 295 (2003).

[164] O'Neil and Gunn, *supra* note 164; See also, Kermisch and Depaus, *supra* note 164; Mepham, *supra* note 164; Jensen et al., *supra* note 164; Schroeder and Palmer, *supra* note 164.

[165] [anon.]

[166] To err on the side of minimizing false positives would create better algorithmic outcomes for people who are unintended recipients, or false negatives, creating worse outcomes for people who were the intended recipients, O'Neil and Gunn, *supra* note 164.

[167] O'Neil and Gunn, *supra* note 164.

[168] Consider some arguments that in the context of algorithms, the "law must make the necessary, yet uncomfortable, shift to outcome-focused analysis", Talia Gillis, *Discriminating Credit Algorithms*, OXFORD BUS. L. BLOG (2020), https://blogs.law.ox.ac.uk/business-law-blog/blog/2020/06/discriminating-credit-algorithms (last visited Oct 8, 2022); Gillis, *supra* note 20; Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017) (noting "the focus on outcomes rather than how an algorithm operates seems especially useful as algorithms become increasingly complicated, even able to modify themselves.").

[169] John Tasioulas, *Artificial Intelligence, Humanistic Ethics*, 151 DÆDALUS, JOURNAL OF THE AMERICAN ACADEMY OF ARTS & SCIENCES 232 (2022); Iason Gabriel, *Towards a Theory of Justice for Artificial Intelligence*, 151 DÆDALUS, JOURNAL OF THE AMERICAN ACADEMY OF ARTS & SCIENCES 218 (2022); O'Neil and Gunn, *supra* note 164.

implementation, a principle that forms the backbone of the Ethical Fairness framework the paper aims to describe.

C. Explainable Fairness Framework: Step-by-step

The Explainable Fairness framework is the process represented in the following flowchart.

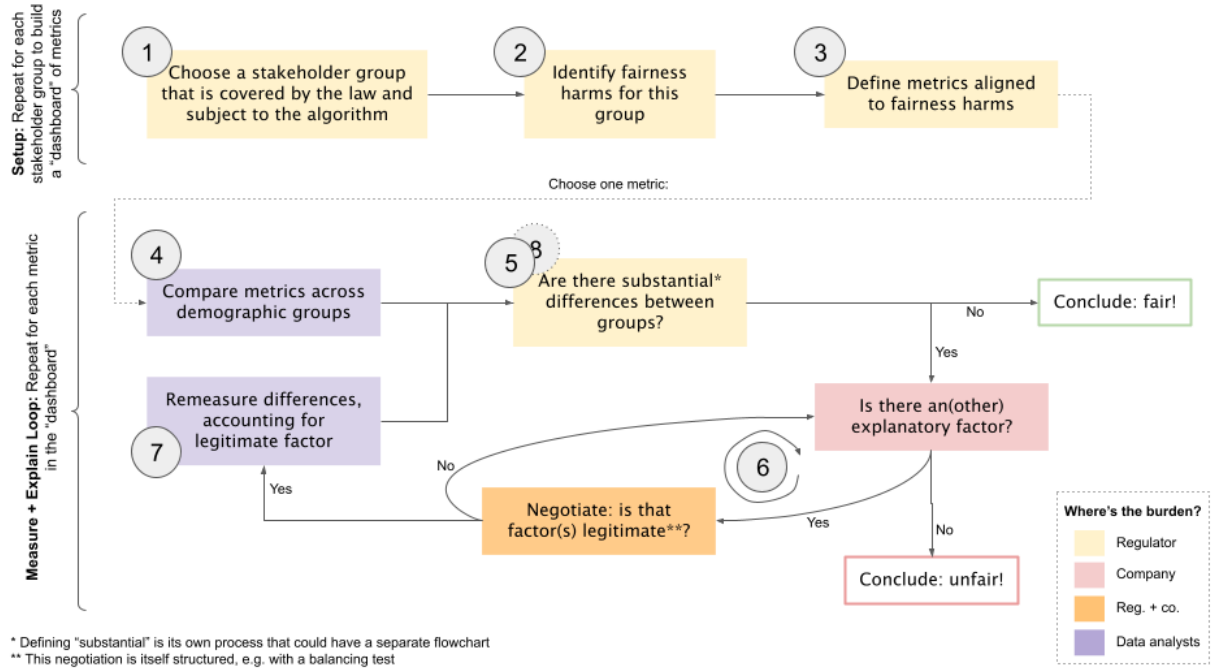

Figure 1. Explainable Fairness flowchart

This section explains the steps of the Explainable Fairness framework, by working through an example of a regulator investigating compliance, with respect to an anti-discrimination law, of an algorithmic system that issues individuals a "risk score".

1. Choose a stakeholder group that is covered by the law and subject to the algorithmic system

During the setup phase of the Explainable Fairness framework (Steps 1-3), regulators create a dashboard of metrics that will be used to measure algorithmic compliance. This is conducted on a per-stakeholder group basis, with each group identified based on legal protections relevant to the algorithm's risk scores. For instance, if a particular anti-discrimination law prohibits sex or race discrimination, the corresponding groups become stakeholders. It is possible for an individual to belong to multiple stakeholder groups.

An area of potential oversight arises when people are covered by anti-discrimination law but are not subject to the algorithm's risk scores.[170] For example, consumers with limited credit history might undergo a different assessment process and not receive a risk score from the same algorithm.[171] This could obscure insights into whether these "thin file" applicants are treated fairly when regulators compare risk scores. Therefore, regulators need to consider how to address individuals who fall outside the algorithmic system's purview but are within the scope of the law.[172] It's important to note, however, that the Explainable Fairness framework is primarily focused on the law's implications for the algorithmic system, rather than offering solutions for such cases.

2. Identify fairness harms for this stakeholder group

The regulator makes a list of harms by constructing an Ethical Matrix as described in Section B. The goal is to identify unreasonable harms against a particular group of people.[173] For instance, a regulator may be concerned about sex discrimination if women are given higher risk scores than men on average, or only if women are more likely than men to be declined for the relevant product

---

[170] *See, e.g.,* Dothard v. Rawlinson, 433 U.S. 321, 330 (1977) (noting that "[t]here is no requirement … that a statistical showing of disproportionate impact must always be based on analysis of the characteristics of actual applicants" in part because "[t]he application process might itself not adequately reflect the actual potential applicant pool, since otherwise qualified people might be discouraged from applying because of a self-recognized inability to meet the very standards challenged as being discriminatory").

[171] BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM, *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit*, 304 (2007); Holli Sargeant, *Algorithmic decision-making in financial services: economic and normative outcomes in consumer credit*, AI ETHICS (2022), https://doi.org/10.1007/s43681-022-00236-7.

[172] This will be familiar to investigating agencies as disparate impact analysis requires the investigator to "take into account the correct population base and it's racial makeup" such that the use of statistical evidence is based on comparison groups that do not extend beyond the total group to which the policy was applied, *see* Darensburg v. Metropolitan Transp. Com'n, 636 F.3d 511, 520 (9th Cir. 2011); Betsey v. Turtle Creek Assoc., 736 F.2d 983, 987 (4th Cir. 1984).

[173] Bryan v. Koch, 627 F.2d 612, 617 (2d Cir. 1980) (the investigating agency must determine whether the alleged consequences are sufficiently adverse or harmful).

or service for which risk scores are part of determining access. Investigating agencies should employ a broad definition of harm and gather all information and evidence available.[174]

3. Define metrics aligned to fairness harms

After identifying potential harms, the regulator must devise one or more corresponding metrics to measure the occurrence and extent of each harm, which will serve as compliance indicators. The defined metrics should accurately reflect changes in the harm scenario; if harm worsens, the metric should consistently increase (or decrease). Metrics should not be influenced by factors other than changes in the algorithm's benefits or harms. Practicality is another key consideration for regulators. Optimal metrics are based on data that is readily available, uniformly collected, and reliable, which often aligns with data gathered for other regulatory requirements such as standard disclosures or reports.

One commonly utilized compliance metric is the four-fifths rule, an impact ratio used in employment discrimination.[175] For a group within a legally protected class (e.g., women in the sex class), the impact ratio is the selection rate for the group divided by the selection rate for the most-selected group within the class.[176] These ratios range from 0 to 1, and a smaller ratio suggests greater disparity in selection rates.[177] Reliable selection data is often available in the private sector as companies keep track of consumer or recruitment interactions, allowing for the calculation of impact ratios.[178]

---

[174] U.S. Dep't of Just., Title VI Legal Manual, § VII: Proving Discrimination — Disparate Impact, at C.1.b. (2021), https://www.justice.gov/crt/fcs/T6Manual7#F ("Agency Practice Tip: While establishing adversity in most cases presents a low bar, investigating agencies nevertheless should employ a broad definition of adversity/harm, and gather any and all evidence of adversity/harm or risk of adversity/harm, including anecdotal evidence from complaining witnesses. Even though such additional evidence may not be required as a legal matter, it provides important context for the decision–maker. Such evidence also informs development of the appropriate remedy in the case of noncompliance.")

[175] *See* Uniform Guidelines on Employee Selection Procedures, 43 Fed. Reg. 38290, 38294 (Aug. 25, 1978) (to be codified at 41 C.F.R. pt. 60-3); Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11996 (Mar. 2, 1979) (to be codified at 29 CFR Part 1607); DEP'T OF LABOR, PRACTICAL SIGNIFICANCE IN EEO ANALYSIS FREQUENTLY ASKED QUESTIONS (last updated Jan. 15, 2021), https://www.dol.gov/agencies/ofccp/faqs/practical-significance.

[176] *See* Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11996 (Mar. 2, 1979) (to be codified at 29 CFR pt. 1607); U.S. Equal Employment Opportunity Commission, *supra* note 70.

[177] *See* Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11996 (Mar. 2, 1979) (to be codified at 29 CFR pt. 1607); U.S. Equal Employment Opportunity Commission, *supra* note 70.

[178] The bias audits required of AI hiring tools by NYC Local Law 144 include calculation of impact ratios for various protected classes, *see* tit. 6, ch. 5 § 5-300.

4. Compare metrics across demographic groups

Next, these metrics are compared across demographic groups, with regulators calculating and comparing each metric defined in Step 3 for relevant groups. For an anti-discrimination law, this might involve comparing between genders, races, ethnicities, age groups, or disability statuses.[179] However, acquiring the relevant demographic data can be a challenge, especially as some companies avoid collecting such information in contexts where anti-discrimination laws apply.[180] In such cases, regulators can either mandate data collection, as done under the Home Mortgage Disclosure Act,[181] or work around missing data, perhaps inferring demographic details from US Census data as the CFPB does in fair lending investigations.[182]

Ultimately, the Explainable Fairness framework requires individual-level demographic information, self-reported or inferred, related to the categories under bias investigation. Once that information is acquired, the essence of this step is looking at outcomes across groups. This could involve comparing means or other distributional statistics, looking at histograms, or doing statistical hypothesis tests or regression analyses.[183]

5. Identify substantial differences between groups in the metrics

In Step 5 of the process, the meaning of any disparities identified in Step 4's comparison of demographic group metrics is examined. If the differences between groups within a protected class are negligible, regulators may conclude there's no discriminatory effect warranting further investigation.

However, when a disparity exists, defining what constitutes a *substantial* difference becomes crucial. Regulators can iteratively define substantial relative to its industry context.[184] If a

---

[179] Regulators could also consider intersectional groups defined by multiple protected classes, e.g., Black women vs White men. The NYC Local Law 144 requires analysis comparing metrics across such intersectional race-gender groups, *see Id.*

[180] Ironically, proponents of this practice sometimes claim "fairness through unawareness": if we do not know applicants' races, then we couldn't possibly discriminate by race. This spurious argument has been thoroughly debunked. Algorithms can produce disparate outcomes by race even without access to any individual's race, because so many other individual characteristics are correlated with race, *see* Gillis and Spiess, supra note 19; Kroll et al., supra note 24 (explaining that a "commonly understood way to demonstrate that a decision process is independent of sensitive attributes is to preclude the use of those sensitive attributes from consideration" is "naive").

[181] Home Mortgage Disclosure Act, Pub. L. No. 94-200, 89 Stat. 1125 (1975) (codified as amended at 12 U.S.C. §§ 2801–2811); Consumer Financial Protection Bureau Notice on Status of New Uniform Residential Loan Application and Collection of Expanded Home
Mortgage Disclosure Act Information about Ethnicity and Race (Sept. 23, 2016), https://files.consumerfinance.gov/f/documents/092016_cfpb_HMDAEthinicityRace.pdf.

[182] *See* CONSUMER FINANCIAL PROTECTION BUREAU, USING PUBLICLY AVAILABLE INFORMATION TO PROXY FOR UNIDENTIFIED RACE AND ETHNICITY (2014), https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/.

[183] RACHEL SCHUTT & CATHY O'NEIL, DOING DATA SCIENCE 34 (2013).

[184] *See infra* Section III.A(2).

substantial difference is identified, the regulator presents the findings to the regulated entity, shifting the burden to the entity to justify the different outcomes under the algorithmic system.

The entity may argue for a legitimate cause for the disparity, suggesting that group-specific characteristics that are considered in the assessment process may account for different outcomes. Once such an explanation is provided, the onus is on the regulator to establish which characteristics can reasonably be used to justify outcome differences between groups and to define the allowable extent of such differences. This step requires the regulator to define legitimate factors.

6. Establish what a legitimate factor is in this context

The regulator's responsibility is to discern which factors are deemed legitimate and which are not. The decision could be based on existing laws, precedent, regulatory guidance, or the results of the quantitative balancing test and causal analysis conducted earlier.[185] This task involves applying the established threshold metric and conducting a causal analysis to evaluate the relationship between the proposed factors and the objective of the algorithmic system.

7. Re-measure the outcome metrics, accounting for legitimate factors

Now, the Explainable Fairness steps require the regulator to incorporate legitimate factors into the analysis previously focused solely on demographic group comparisons. The crucial question is whether the disparities between demographic groups diminish once legitimate factors are factored in.

Conceptually, the regulator will now consider whether a risk score is higher for impacted individuals of a certain demographic after accounting for a legitimate factor. Various techniques could be used for this, like including the factor as an independent variable in a regression, or stratifying the data based on it. The framework doesn't endorse a specific approach but supports the concept of allowing legitimate factors to mitigate observed disparities between groups.

8. Identify persistent substantial differences in the metrics

When substantial disparities between groups are eliminated after factoring in legitimate factors, the regulator may conclude that while demographic differences exist, they can be justified by these factors, eliminating the need for further investigation.

Conversely, if significant disparities persist even after considering all legitimate factors, the regulator might conclude that discrimination is at play. In this case, based on their jurisdiction, they could utilize evidence gathered through the Explainable Fairness framework to determine the

[185] *See infra* Section III.A(3).

most effective enforcement action to prevent the continued use of unfair or unlawful algorithmic systems.

IV. Use Cases

This Article presents two hypothetical examples to show how the Explainable Fairness framework can be applied.

A. Student lending

The application of the Explainable Fairness framework is well illustrated by positing a state financial services regulator that aims to audit a student loan underwriting algorithm. Suppose the regulator is concerned with race-based discrimination in breach of state fair lending laws.[186] And suppose that the regulated entities (i.e., lenders) are using algorithms in their underwriting processes that take as input a completed loan application and output a detailed loan offer for that applicant. In these circumstances, the following steps may guide the regulator's audit.

1. Consider the population subject to this algorithm

For simplicity, suppose all student loan applications are input to the one algorithm. Thus, the regulator is not missing any affected persons by focusing on the one algorithmic underwriting system.

2. Identify fairness harms

The regulator can declare what kinds of harms are the focus of its investigation of this algorithm. It could do so by constructing an Ethical Matrix described in Section III.B. For the purpose of this Article, the regulator may decide that in the context of student loans, it is concerned about the risk of discrimination, specifically if a certain racial group of applicants is (1) less likely to be approved for a loan, (2) pays a higher interest rate, or (3) refusal to grant in substantially the amount requested in an application.

[186] This analysis focuses on a hypothetical regulator investigating potential discriminatory practices in algorithmic student lending. This Article does not interrogate the legal framework governing discrimination in credit pricing. Beyond other federal laws, there are many state and local laws with discrimination provisions, such as fair housing laws.

The regulator should employ a broad definition of harm and gather any information and evidence available. Consulting with affected persons, consumer advocates or other domain experts could be helpful in identifying and defining the potential harms.[187]

3. Define metrics aligned to fairness harms

In order to translate fairness harms into metrics, the regulator must declare ways of measuring whether, and to what degree, the harm is occurring. In the case of student loans, it may declare the following outcomes for qualitative analysis:

1. Approval rate: the proportion of applicants that are granted a loan compared to those declined a loan.
2. Annual Percentage Rate (APR): the cost of credit expressed as a yearly rate in a percentage.[188]
3. Granted loan amount compared to requested amount: defined as the amount of the loan offer, divided by the amount requested in an application.

These metrics are practical and convenient as information that is routinely and uniformly collected by lenders and regulators.[189]

4. Compare metrics across demographic groups

Assume the state regulator has reliable and detailed loan application data from all applicable lenders in the state, sufficient to analyze these outcomes by metrics above. Since lenders do not collect race information about consumers,[190] the regulator may use Bayesian Improved Surname Geocoding (BIFSG) to infer the race of applicants for the analysis.[191]

That would require each lender to submit a spreadsheet with one row of data per loan application received in the prior calendar year.[192] Each row of data would include the following columns:

---

[187] *See infra* Section III.B.
[188] Congress passed the Truth in Lending Act in 1968 which required lenders to use uniform annual percentage rate (APR) terminology, Pub. L. No. 90-321, 82 Stat. 146 (1968) (codified in 15 U.S.C. §§ 1601-1667e).
[189] *See, e.g.*, Student Banking Reports to Congress, CFPB (2022), https://www.consumerfinance.gov/data-research/student-banking/student-banking-reports-congress/.
[190] ECOA and Regulation B generally prohibit a creditor from inquiring "about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction", 12 C.F.R. § 1002.5(b).
[191] This race inference methodology leverages US Census data and is widely used, including by CFPB. See Ioan Voicu, *Using First Name Information to Improve Race and Ethnicity Classification*, 5 STATS. & PUB. POL'Y 1 (2018); CONSUMER FINANCIAL PROTECTION BUREAU, *Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment*, (2014), http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.
[192] There is nothing special about a year; the regulator could request a different time horizon. In general, the more data the regulator gets, the better chance they have to detect even subtle differences between groups in outcomes.

- First name, surname, and address of applicant (used for BIFSG inference);
- Loan amount requested;
- Underwriting decision (Approved/Declined);
- Loan amount offered; and
- APR offered.

This information then gives the regulator the necessary data to create demographic groups (via inference) and calculate the average value of each metric from Step 3 for each protected group. For a given group:

- Average approval rate would be the number of applicants Approved divided by the total number of applicants (number Approved + number Denied) in the group.
- Average APR would be the mean of "APR offered" across all applicants in the group who were Approved.
- Granted loan amount compared to requested amount would be: For each applicant in the group who was Approved, calculate "offer ratio" as "loan amount offered" divided by "loan amount requested". Then calculate the mean of "offer ratio" across all applicants in the group who were Approved.

This example focuses on the outcome metric of approval rate.

5. Identify substantial differences across groups in the metrics

Suppose then that the regulator finds, for a specific lender, the average approval rate was 70% among inferred-Black applicants, and 92% among inferred-White applicants. The regulator must then define *substantial* by reference to the industry. Suppose in this instance, the regulator draws inspiration from the EEOC four-fifths rule,[193] and the regulator decides this identified difference in metrics across racial groups is big enough to warrant further investigation, since 70% is less than four-fifths of 92%. The regulator shows its analysis to the lender.

6. Establish what a legitimate factor is in this context

The lender may respond that there are legitimate factors explaining why the inferred-Black applicants are declined at higher rates compared with inferred-White applicants. For example, the lender may argue such applicants have lower FICO scores, attend lower-ranked colleges, and have declared majors that tend to be less lucrative, all factors that are commonly considered by lenders.

---

[193] *See* Uniform Guidelines on Employee Selection Procedures, 43 Fed. Reg. 38290, 38294 (Aug. 25, 1978) (to be codified at 41 C.F.R. pt. 60-3); Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11996 (Mar. 2, 1979) (to be codified at 29 CFR pt. 1607); U.S. DEP'T OF LABOR, PRACTICAL SIGNIFICANCE IN EEO ANALYSIS FREQUENTLY ASKED QUESTIONS (last updated Jan. 15, 2021), https://www.dol.gov/agencies/ofccp/faqs/practical-significance.

The lender may argue that it has observed that these factors are predictive of default and therefore legitimate to consider in underwriting.

Now the regulator must determine whether it agrees that each of the lender's proposed factors is legitimate. At the heart of this question is a tradeoff in lending markets. On one hand, accurately analyzing creditworthiness is important to a bank's profitability, exposure to credit risk, and portfolio quality; it is similarly important for consumers that they are offered access to credit, but only to the extent they can repay.[194] Ensuring responsible lending is essential to protect consumers from over-indebtedness and potential bankruptcy.[195] Conversely, if a given characteristic is highly correlated with an applicant's race, then if used in underwriting it could lead to disparities across races, which is unfair and potentially unlawful. As explained in Section III.C., while there is a lack of clarity on the issue of proxy variables, recent scholarship highlights the normative significance and challenge of proxy discrimination.[196]

In the absence of a clear definition and approach to anti-discrimination procedures related to proxy variables, it will be challenging to translate moral assumptions about legitimate factors because of the unique use of proxy variables in algorithmic contexts. So, below is an example of the "ratio test" that can be used to navigate this tradeoff.[197] For a given characteristic F:

- How well does F predict default? This is a data analytics question to be answered with regression analysis or a similar procedure.[198] The numeric answer is the numerator of the ratio test.

- To what extent is F a proxy for protected class? This is also a data analytics question to be answered with a correlation analysis or a similar procedure. The numeric answer is the denominator of the ratio test. The regulator may set a strict maximum value for the denominator, so that very strong proxies for race are never declared legitimate.

---

[194] See discussion in Sargeant, *supra* note 171; Barbara Kiviat, *The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores*, 84 AMERICAN SOCIOLOGICAL REVIEW 1134 (2019); FREDERIC MISHKIN, THE ECONOMICS OF MONEY, BANKING, AND FINANCIAL MARKETS (12 ed. 2019); KENT MATTHEWS & JOHN THOMPSON, THE ECONOMICS OF BANKING (2 ed. 2008).

[195] For instance, US Consumer Reports highlights that it is still very difficult for US students to calculate the true cost of student loans and many are becoming heavily indebted, Consumer Reports, *Student Loan Debt Crisis*, https://www.consumerreports.org/student-loan-debt-crisis/ (last visited Jun. 20, 2023); Jami Hubbard-Solli, *Responsible lending: An international landscape*, CONSUMERS INTERNATIONAL (2013), https://www.consumersinternational.org/media/2246/ciresponsiblelending_finalreport_06-11-13.pdf.

[196] Schwarcz and Prince, *supra* note 153; Hellman, *supra* note 152 (referencing discussion of proxy variables in oral arguments in Students for Fair Admissions, Inc. v. The President and Fellows of Harvard College, No. 20-1199 and Students for Fair Admissions, Inc. v. University of North Carolina, No. 21-707 (Oct. 31, 2022)).

[197] *See infra* Appendix.

[198] SCHUTT AND O'NEIL, *supra* note 184, at 34.

- If the ratio (numerator divided by denominator) exceeds the threshold value chosen by the regulator, then the factor in question is declared legitimate.[199] The regulator's justification for this is that the factor's value in predicting the important outcome of default "outweighs" its role as a proxy for race.

This ratio test can be used with multiple factors. There are some special considerations in this case. Since the order of legitimate factors should not matter, this process evaluates how well several factors *together* predict risk and are a proxy for protected class, which is a more complicated question, with more modeling choices to make. The regulator may also want to set a strict maximum value for the extent to which the entire set of legitimate factors are jointly a proxy for race. This requirement is intended to avoid a scenario where a set of factors are declared legitimate that are individually innocuous but, taken together, are a strong proxy for race. These issues are discussed further in the Appendix.

One example of regulator discretion may be that FICO score passes the ratio test, but college rank and student's major do not. The regulator may also conduct a causal analysis that considers the legitimacy of the plausible story between FICO score and creditworthiness.[200] Then, FICO score is declared the sole "legitimate factor" to explain differences in approval rates of a lender.

7. Re-measure the outcome metrics, accounting for legitimate factors

Now the regulator may request the FICO score of each applicant included in the analysis. This can be thought of as requesting another column of the dataset described in Step 4.

With this augmented dataset, the regulator then re-measures approval rate by inferred-race group, now controlling for FICO score. Technically there are many ways to "control for" FICO. Bins of FICO scores could be created (e.g., 650-700, 701-750, 751-800, and so on) and then *within each bin* comparing the approval rate between inferred-race groups. These segmented analyses answer the narrower question of whether among applicants with similar FICO scores that are at a similar risk of default, do approval rates differ by inferred race. Note that this phrasing of the question incorporates a context-specific notion of a similar applicant, and this notion arose from the deliberate approach to legitimate factors outlined in the previous step.

8. Identify persistent substantial differences in the metrics

If the regulator does this analysis and finds there is now not a large difference between race groups in approval rates within any bin of FICO score. Then the regulator can then conclude that the

---

[199] *See infra* Appendix.
[200] *See* Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 194 (2017); Citron and Pasquale, *supra* note 8.

overall race gap in approval rates between race groups was explained by differences in FICO scores.

Therefore, on this specific harm and metric, there is no need for further enquiry. However, a regulator may exercise their discretion and continue to repeat such analysis across various harms and metrics in the performance of its duties.

B. Short-term Disability Insurance

A second illustrative example is set out below to further develop the practical implementation of the Explainable Fairness framework.

Consider the use of an algorithmic system in short-term disability insurance. In this example, employers purchase short-term disability insurance coverage for their workers. A covered employee can submit a claim if they have an injury or condition that temporarily keeps them from working. If the insurer approves a claim, it sets a time limit, 10 weeks for example. This means the insurer will effectively pay the employee to stay home for 10 weeks, after which the employee should return to work; or, if a doctor decides at that point they need longer to recover, then the claim may be extended.

Suppose some insurers use algorithms to make preliminary approvals of claims. When a new claim is received, it is input into an algorithm which can approve a claim based on the claim details (e.g., description of the injury or condition, clinical notes from related doctor visits) as well as information about the person's policy (e.g., what kinds of coverage do they have, what are the limits). If the algorithm approves the claim, the review is over, and the claim gets paid. If the claim is rejected, the claim is sent to a human decision-maker at the insurer, who can approve or deny it. The insurer's *human* decision is final.

Many US states have rules prohibiting unfair discrimination in a variety of insurance practices including the handling of claims.[201] In particular, these rules imply that similar claims may not be handled differently on the basis of the claimant's race, ethnicity, or other protected-class status. For this example, suppose a state insurance commissioner (i.e., the regulator) wants to investigate whether the algorithmic systems just described meet this standard.

1. Consider the population subject to this algorithm

Assuming all short-term disability claims go through one algorithmic system, no specific groups of policyholders are left out of the regulator's analysis.

---

[201] *See infra* Section II.B.(3).

2. Identify fairness harms

Assume the regulator's principal concern is potential unjust discrimination in claim handling. Crucially, the claims handling process transcends the preliminary approval algorithm, necessitating a broader view of outcome disparities. For instance, if White claimants receive more preliminary algorithmic approvals than Black claimants, this may be construed as unfair discrimination.

However, if human adjusters balance the scales by approving Black claimants at a higher rate, leading to equal overall approval, the regulator must consider whether the algorithmic discrepancy is still discriminatory. Alternatively, if initial approval rates are equivalent for both Black and White claimants from the algorithm and adjusters, but White claimants are more likely to secure extensions, resulting in longer recovery periods, a methodical approach is needed to determine if this constitutes unfair discrimination. At least in this example, it would warrant inclusion in an Ethical Matrix as a potential failure of the algorithm for Black claimants.

Identifying harms, specifically defining what constitutes unfair discrimination in claims handling, is a value-laden question rather than a technical one. It should be assessed not just by the algorithm's outputs, but in the broader context of the entire system that generates outcomes for stakeholders. This comprehensive view would include the entire claim handling process which determines a person's recovery period following an injury or other condition.

3. Define metrics aligned to fairness harms

Translating these concerns into specific measurements, the regulator may declare that the outcomes of interest for quantitative analysis are:

1. Approval rate for short term disability claims.
2. Number of days approved for approved claims (excluding any extensions).
3. Number of extensions sought and approved.
4. Total days covered for approved claims (including all extensions).

4. Compare metrics across demographic groups

The regulator may request claims data from the insurers to calculate these metrics. As in the previous example, suppose insurers do not collect race information about consumers, so the regulator will use BIFSG to infer the race of applicants for the analysis.[202]

Each insurer submits a spreadsheet with one row of data per short-term disability claim received within a given time window. Each row of data includes the following columns:

- First name, surname, and address of claimant (used for BIFSG inference);
- Claim duration requested (days);
- Claim duration approved (days; blank if initial claim was not approved);
- Number of extensions requested for this claim (blank if initial claim was not approved);
- Number of extensions approved for this claim (blank if initial claim was not approved); and
- Total duration covered (days; include initial claim and all extensions).

This gives the regulator what it needs to create demographic groups (via inference) and calculate each metric for each group. The rest of this example will focus on the metric total duration covered.

5. Identify substantial differences across groups in the metrics

Suppose the regulator finds, for a certain insurer, the average total days covered is 60 for White claimants but only 45 for Black claimants, and the difference is highly statistically significant. That is, White claimants are getting a better outcome – an extra 15 days paid time off work to recover from their injuries – than Black claimants. The regulator brings this finding to the insurer.

6. Establish what a legitimate factor is in this context

The insurer justifies the greater total days covered for White claimants, arguing that they are older, have more comorbidities, and their claim-causing injuries/incidents were different. Additionally, the insurer notes that White claimants had higher FICO scores, which they claim generally correlate with total claim duration. Therefore, the insurer suggests that the credit score disparity between White and non-White claimants also explains the racial gap in total claim duration.

To validate these factors, the regulator applies a ratio test to claimant age and comorbidities to evaluate their legitimacy in explaining total claim length disparities. As older people and those

[202] *See* Voicu, *supra* note 192; CONSUMER FINANCIAL PROTECTION BUREAU, *supra* note 192.

with comorbidities typically require longer recovery periods, these factors passing the ratio test appear logical.[203]

In the earlier context of student lending, FICO scores were considered a legitimate factor due to their design as a creditworthiness measure.[204] However, this argument is challenged in the context of short-term disability insurance claims handling. In this scenario, the regulator argues that FICO scores can contribute to discriminatory practices and dismisses it as an illegitimate factor. The goal here is to assess the treatment of similar claims without factoring in creditworthiness. The legitimacy of factors varies contextually. While a factor may be legitimate in one setting, it might not be in others. For example, several states ban or limit the use of credit scores in determining policy rates in certain circumstances.[205]

In Washington, Insurance Commissioner Mike Kreidler tried to install a three-year ban on insurers using credit information to set auto, homeowner, and renter insurance.[206] However, the ban was stayed due to the rule exceeding the Insurance Commissioner's authority.[207] The Court did nonetheless accept that "there is an undeniable link between race and poverty, and any policy that discriminates based on creditworthiness correspondingly results in a disparate impact on communities of color. The temporary rule does in fact protect from such discrimination".[208] Therefore, in such a context, FICO scores might not be deemed a legitimate factor because the regulator may view the connection between race and poverty as untenable.

---

[203] *See* Sohail M. Mulla et al., *Factors Associated with the Duration of Disability Benefits Claims among Canadian Workers: A Retrospective Cohort Study*, 5 CMAJ OPEN 109 (2017) (showing that for "both short- (n = 70 776) and long-term disability (n = 22 205) claims, and across all disorders, older age, female sex, heavy job demands, presence of comorbidity, attending an independent medical evaluation, receipt of rehabilitation therapy and longer time to claim approval were associated with longer claim duration.").

[204] *See* Kiviat, *supra* note 195.

[205] *Credit-Based Insurance Scores*, NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS, https://content.naic.org/cipr-topics/credit-based-insurance-scores (last visited Apr. 27, 2023) (explaining California, Hawaii, Maryland, Michigan, and Massachusetts ban or limit insurance companies' use of credit scores, Oregon and Utah have limited prohibitions. Insurers argue that the use of credit-based insurance scores is necessary to properly evaluate risk. However, consumer groups and many states accept that the use of credit-based insurance scores falls disproportionately on certain minority and low-income groups).

[206] Washington State Office of the Insurance Commissioner, *Kreidler Adopts Rule Temporarily Banning Credit Scoring, Proposes Rule to Increase Transparency* (Feb. 1, 2022), https://www.insurance.wa.gov/news/kreidler-adopts-rule-temporarily-banning-credit-scoring-proposes-rule-increase-transparency (last visited Apr. 27, 2023).

[207] NAMIC et al. v. Office of the Ins. Comm'r of the State of Wash., No. 22-2-00180-34 (Wash. Super. Ct. Aug. 29, 2022).

[208] Transcript of Oral Decision at 45, NAMIC et al. v. Office of the Ins. Comm'r of the State of Wash., No. 22-2-00180-34 (Wash. Super. Ct. July 29, 2022).

This decision to disregard FICO scores did not rely on the ratio test or any quantitative assessment. It stems from the regulator's judgment that creditworthiness should not influence the similar treatment of similar claims.

7. Re-measure the outcome metrics, accounting for legitimate factors

Now the regulator requests augmented data incorporating the legitimate factors for analysis, effectively expanding the dataset to include claimant age and comorbidities.[209]

With this augmented data, the regulator focuses on determining whether disparities exist between inferred-White and inferred-non-White claimants in total claim duration, while controlling for claimant age and comorbidities.

8. Identify persistent substantial differences in the metrics

In this scenario, racial disparities persist even after accounting for legitimate factors. The regulator then presents these results to the insurer who may propose additional factors to account for the difference, thereby revisiting Step 6. This process iterates until the difference is either fully attributed to legitimate factors, or the insurer ceases to propose further explanatory factors.

V. Conclusion

The complexity and opacity of algorithms, combined with their near omnipresence in what historically were human bureaucracies, presents society with a dilemma: either allow algorithmic discrimination to run rampant and unchallenged, or develop a way to enforce context-specific anti-discrimination law that can be tested, monitored, and modified over time.

Regulators need an approach to defining fairness that takes into account precedent, principle, and practicality, that is understandable to most or all stakeholders, and that is open to feedback, criticism, and legal appeal.

The Explainable Fairness framework takes on this challenge. It is an approach in which the definition of fairness is an ongoing and open negotiation between the deployers of algorithms and the regulators in charge. In particular, the technical expertise required to build an algorithm is not required to regulate it; instead, only the legal notions of what differentiates a legal and legitimate method of discrimination versus an illegal and illegitimate approach is necessary, as well as the

[209] Note that there are important choices to make regarding the data structure of the comorbidities columns. For instance, does every comorbidity get its own "code" and get considered separately? Or are there groups of comorbidities that are similar to each other? Or is there a "comorbidities index" that puts all comorbidities, and combinations of comorbidities, on a common numeric scale? Each alternative has benefits and drawbacks, and the regulator should get a data science expert to lay them out.

ultimate decision on when a difference in outcomes constitutes a *substantial* difference. Even there, the Explainable Fairness framework offers a positive feedback loop to encourage tighter definitions of fairness over time.

Explainable Fairness does not require the underlying algorithmic system to be explainable or explained. This is both a strength and a weakness; a strength because it can be applied to any algorithm with any amount of complexity, a weakness because, when completed, the process does not explain the reasoning behind the algorithm itself, only the reasoning for why the process has been deemed fair. The good news is that this reasoning should be understandable to all of the stakeholders.

In high-stakes algorithmic contexts such as hiring, insurance, credit, or housing decisions, consumers are less concerned with the exact workings of the algorithms and more interested in whether individuals in their demographic are being treated fairly. There is precedent for this concern and a growing need for a robust solution. This solution could take the form of regulatory enforcement, incorporating a clearly defined procedure like the Explainable Fairness framework. This is the ultimate product of the Explainable Fairness framework and will be a useful paradigm for many use cases where regulators are currently lacking a viable approach.

VI. Appendix: Ratio test

We propose a ratio test as one way a regulator could decide whether a given set of factors is legitimate, in the sense that a regulated company can use the factors to "explain away" differences between protected-class groups in a given outcome of interest. In this appendix we will use the example of race/ethnicity groups as a protected class.

Structure of the ratio

The basic idea behind the ratio is to compare two aspects about a given set of factors proposed:

(1) The factors are predictive of the risk (or value) the consumer represents. They provide consumer-level information that helps the regulated company make the *right* decision about that consumer (in a context-specific sense that the regulator can articulate); and

(2) The factors encode some information about consumers' race/ethnicity. The extent to which they do this undermines their value as "explainers" of race differences. In an extreme case, where a proposed factor is an exact proxy for race, for that factor to "explain away" a race gap is tautological. The numerator of the ratio will be a numeric measurement of (1), and the denominator will be a numeric measurement of (2).

Motivating Example: One legitimate factor

As a simple example, imagine the context is credit underwriting, the proposed factor is FICO score, and race is modeled as binary (White=1 vs non-White=0). In this case the numerator could be the correlation coefficient between FICO score and (binary) default, and the denominator could be the correlation coefficient between FICO score and (binary) race.

Modeling Choices for multiple factors

Here are some considerations and examples of specific measurements that can be used:

(1): "Predictive of the risk (or value) the consumer represents"

- This begs the question: how is risk (or value) defined? It depends on the regulated activity being considered. In credit, for instance, an underwriter's job is to assess the risk that the applicant will default, and decide accordingly whether to offer a loan and at what interest rate. In this context, "risk" is "risk of default", and the ratio test numerator could address the question, "How predictive are these factors of default?" There is room for debate here. An underwriter might argue that they are not assessing the risk of default, but estimating what percentage of the total debt the borrower will ultimately pay; this would imply a different analysis ("How predictive are these factors of full payment?").

- Suppose a consumer's risk (or value) has been defined with reference to some observable outcome X (e.g., "default" or "percentage of total debt paid"), so we can ask the question: "How predictive of X are the proposed factors?" A numeric answer to this question is a goodness-of-fit statistic.
    - In one-factor scenarios like the FICO example above, the correlation coefficient between the factor and X can be an option for the numerator.
    - Linear regression can be used to model X as a function of the proposed factors. In this case the R^2 value from the regression model would be an option for the numerator.
    - Any technique could be created to predict X based on the proposed factors, then use a general fit statistic like accuracy or F1 score.

(2): "Encode some information about consumers' race/ethnicity"

- This is also approached as a prediction problem. To what extent do the proposed factors predict race? As above, a goodness-of-fit statistic offers a numeric answer.
- Since race/ethnicity is categorical (not continuous) with non-ordered categories, classification models might be better for predictions than simple linear regressions. This suggests goodness-of-fit statistics like accuracy or F1 score.
- There are consequential choices to make about the structure of race/ethnicity in the data. Is each race its own category, or are they aggregated in a binary analysis (e.g. White vs non-White)? Does each distinct combination of races/ethnicities get its own category, or is there a catch-all "Multiracial" category? Is Hispanic/non-Hispanic a separate classification from race that cuts across race categories, or is Hispanic a distinct race?
    - In one-factor scenarios like the FICO example above, if race is modeled as binary then the correlation coefficient between the factor and (binary) race can be an option for the denominator.

<u>Threshold value of the ratio and other constraints</u>

Intuitively, if the ratio is large (i.e., (1) dominates (2)), it suggests the proposed factors are legitimate: together they provide a lot of information about consumers that is relevant to the company's decision, but they do not effectively encode race/ethnicity such that the company could explain away arbitrary race disparities. Conversely, if the ratio is small, it suggests the proposed factors are not legitimate: they say relatively little about the consumer that is relevant to the company's decision, and relatively a lot about the consumer's race.

The numeric range of the ratio depends heavily on the modeling choices just discussed. In practice, once those choices are made, a regulator might want to "calibrate" the ratio by working with sample data. Specifically, the regulator could identify some factors that are patently legitimate (perhaps based on precedent or industry standards) and some factors that are patently not

legitimate, then calculate the ratio for each factor. Presumably the ratio's value for each patently-legitimate factor will be higher than for the patently-illegitimate ones, and an appropriate threshold would be somewhere in the middle.

Ultimately the idea is to specify a threshold value: if the ratio for a given set of proposed factors is above the threshold, then the factors are legitimate; if below, then they are not.

In addition to a threshold value for the ratio overall, a regulator could place separate constraints on its components. There could be a ceiling for the denominator -- i.e., a limit on the extent to which the proposed factors can predict race (no matter how well they can predict risk). There could also be a floor for the numerator -- i.e., a minimum degree of risk-prediction required for a set of factors to be legitimate (no matter how little race/ethnicity information they encode).