Using Algorithms to Tame Discrimination: A Path to Diversity, Equity, and Inclusion

Deven R. Desai,[†*] *Swati Gupta,*[**] *and Jad Salem*

Companies that try to address inequality in employment face a paradox. Failing to address disparities regarding protected classes in a company's workforce can result in legal sanctions; but proactive actions to address and avoid such disparities can also face legal scrutiny and sanctions too. After the summer of 2020, companies such as Microsoft announced large programs to address inequity in employment. They soon received letters from the Labor Department's Office of Federal Contract Compliance Programs ("OFCCP") because of the OFCCP's concern that the plans will end up discriminating based on race. At the same time, the OFCCP announced a settlement with Microsoft on September 19, 2020, for $3 million back pay and interest to address hiring disparities "against Asian applicants" for several positions from December 2015 to November 2018. These examples are not isolated and are likely to persist. Any company seeking to identify talent will likely use data and algorithms to screen and hire employees. That practice will again raise the tension of how to increase diversity without running into problems of embedded inequity and making decisions that are prohibited because they are based on protected class status. We offer a potential path forward to solve this

 This Article has benefitted from discussions with and input from Professors Jason Bent, Pauline Kim, and Peter Lee, and from participants at the Law and Ethics of Big Data Workshop 2020, Privacy Law Scholars Conference 2021 and FAccT 2022. We thank all for their insights.

[*] Associate Professor of Law and Ethics, Georgia Institute of Technology; J.D. Yale Law School; Affiliated Fellow, Yale Law School Information Society Project; former Academic Research Counsel, Google, Inc.

[**] Assistant Professor, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology.

paradox by exploring current advances in Computer Science and Operations Research.

By carefully acknowledging uncertainties in candidates' data (using the framework of partially ordered sets), a hiring entity can improve equal opportunity practices. The solution is to embed error-mitigation due to uncertainties or biases in an algorithmic decision-making process without crossing into illegal discriminatory practices (e.g., without enforcing quotas). In short, this Article explains a way to design fair screening methods that account for biases and uncertainties in data and abide by anti-discrimination law.

TABLE OF CONTENTS

> *Individuals typically act on incomplete information and with subjectively derived models that are frequently erroneous; the information feedback is typically insufficient to correct these subjective models.*[1]
>
> *— Douglass C. North, 1993*

> *[T]he difficulties generated for scientific inquiry generated by unconscious bias and tacit value orientation are rarely overcome by devout resolutions to eliminate bias. They are usually overcome, often only gradually, through the self-corrective mechanisms of science as a social enterprise.*[2]
>
> *— Ernest Nagel, 1961*

INTRODUCTION

How employers identify whom to interview and then hire has important effects across society.[3] Employment significantly affects access to

[1] Douglass C. North, Lecture to the Memory of Alfred Nobel: Economic Performance Through Time (Dec. 9, 1993), https://www.nobelprize.org/prizes/economic-sciences/1993/north/lecture/ [https://perma.cc/KUV3-RTVS].

[2] ERNEST NAGEL, THE STRUCTURE OF SCIENCE: PROBLEMS IN THE LOGIC OF SCIENTIFIC EXPLANATION 489 (1961).

[3] *See, e.g.*, CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2017) (discussing ways software can affect employment conditions); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 34-35 (2015) (describing use of software and online data to make hiring decisions); Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671 (2020) (examining potential bias in hiring algorithms); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016) (outlining how data-driven decision-making in hiring can reflect biases present in data, resulting in discriminatory decisions).

healthcare, continuing education, and quality of life, for employees, their families, and the next generation.[4] The benefits of employment are not, however, evenly distributed across race and gender categories in the United States.[5] This problem leaves employers with the challenge of how to address systemic under-hiring and promotion of under-represented minorities and genders. The challenge is part of pursuing diversity, equity, and inclusion ("DEI"), which has spawned an industry that was estimated at $8 billion annually in 2003[6] and that one 2021 report projects to hit more than $15 billion annually by 2026.[7] Companies thus must solve this problem: How does a hiring entity take DEI aspirations and turn them into real outcomes? By bringing insights from data analytics, new computer science methods to address uncertainty in datasets, and legal rules together, this paper offers an answer.

After George Floyd's death, several companies announced plans to address racial injustice in employment. Microsoft announced a $150 million investment to improve diversity including setting a goal of doubling the number of "Black and African American people managers, senior individual contributors and senior leaders" in the United States by

[4] Ajunwa, *supra* note 3, at 1681 n.40 (noting research showing a correlation between unequal employment opportunities and "violence, incarceration, drug abuse, obesity, teenage pregnancy, and mental health issues").

[5] *See, e.g.*, Pamela Newkirk, *Diversity Has Become a Booming Business. So Where Are the Results?*, TIME (Oct. 10, 2019, 6:10 AM EDT), https://time.com/5696943/diversity-business/ [https://perma.cc/2YT4-3JAH] ("People of color–who make up nearly 40% of the U.S. population–remain acutely underrepresented in most influential fields. From 2009 to 2018 the percentage of [B]lack law partners inched up from 1.7% to 1.8%. From 1985 to 2016, the proportion of [B]lack men in management at U.S. companies with 100 or more employees barely budged–from 3% to 3.2%. People of color held about 16% of Fortune 500 board seats in 2018. A 2018 survey of the 15 largest public fashion and apparel companies found that nonwhites held only 11% of board seats and that nearly three-quarters of company CEOs were white men. And in the top 200 film releases of 2017, minorities accounted for 7.8% of writers, 12.6% of directors and 19.8% of lead roles.").

[6] *See* Fay Hansen, *Diversity's Business Case Doesn't Add up*, WORKFORCE.COM (Apr. 2, 2003), https://workforce.com/news/diversitys-business-case-doesnt-add-up [https://perma.cc/S9CB-BA3T] (quoting MIT Professor Thomas A. Kochan: "There are estimates that companies spend $8 billion on diversity training annually").

[7] Glob. Indus. Analytics, Inc., *With Global Spending Projected to Reach $15.4 Billion by 2026, Diversity, Equity & Inclusion Takes the Lead Role in the Creation of Stronger Businesses*, PR NEWSWIRE (Nov. 3, 2021, 10:50 AM ET), https://www.prnewswire.com/news-releases/with-global-spending-projected-to-reach-15-4-billion-by-2026--diversity-equity--inclusion-takes-the-lead-role-in-the-creation-of-stronger-businesses-301413808.html [https://perma.cc/PJ9L-TVXX] ("The global market for Diversity and Inclusion (D&I) estimated at US$7.5 Billion in the year 2020, is projected to reach a revised size of US$15.4 Billion by 2026 . . . .").

2025.[8] Wells Fargo made a commitment to "double Black Leadership" by 2025 and "will evaluate senior leaders based on their progress in improving diversity and inclusion in their areas of responsibility, in addition to other efforts."[9] Google set a goal of having 30% of its leadership from "underrepresented groups" by 2025.[10] Boeing seeks to increase representation of "Black employees by 20% while boosting other underrepresented groups over the next three years."[11] Adidas announced plans to fill at least "30% of new positions with [B]lack or Latinx people."[12] Although many people may laud these programs, both Microsoft and Wells Fargo received letters from the Labor Department's Office of Federal Contract Compliance Programs ("OFCCP") due to concern that the plans may discriminate based on race.[13] At roughly the same time, the OFCCP announced a settlement with Microsoft in September 2020 for $3 million in back pay and interest to address hiring disparities "against Asian applicants" for several positions from December 2015 to November 2018.[14]

The two OFCCP positions clash and appear to create a paradox: inaction opens a company to litigation, if not breaking the law, and corrective action creates the same risks.[15] In addition, imagine leading a company

---

[8] Clare Duffy, *Plans at Microsoft and Wells Fargo to Increase Black Leadership Are Under Scrutiny from the Labor Dept.*, CNN BUS. (Oct. 7, 2020, 5:49 PM EDT), https://www.cnn.com/2020/10/07/business/microsoft-wells-fargo-diverse-hiring-probe/index.html [https://perma.cc/66KM-UR3F].

[9] *Id.*

[10] Dina Bass & Josh Eidelson, *Microsoft, Wells Fargo Diversity Plans Draw U.S. Labor Inquiry*, BLOOMBERG L. (Oct. 6, 2020, 2:43 PM), https://news.bloomberglaw.com/daily-labor-report/microsoft-plan-to-add-black-executives-draws-u-s-labor-inquiry [https://perma.cc/QLJ3-TTY6].

[11] *Id.*

[12] Clare Duffy, *Adidas Says At Least 30% of New Positions Will Be Filled by Black or Latinx People*, CNN BUS., https://www.cnn.com/2020/06/09/business/adidas-diverse-hiring-initiative/index.html (last updated June 9, 2020, 11:35 PM EDT) [https://perma.cc/3H36-XHYL].

[13] *Id.*

[14] Off. of Fed. Cont. Compliance Programs, *U.S. Department of Labor and Microsoft Corp. Enter Agreement to Resolve Alleged Hiring Discrimination Affecting 1,229 Applicants in Four States*, U.S. DEP'T OF LAB. (Sept. 18, 2020), https://www.dol.gov/newsroom/releases/ofccp/ofccp20200918 [https://perma.cc/AA7R-EU9W].

[15] One might argue that the recent OFCCP inquiries were peculiar to the Trump administration's approach to this area of law and not something the current administration would pursue. Administrations, however, change and a new one might follow the Trump approach. Regardless of who is in the White House, legal activism to challenge steps taken to address diversity or challenge discriminatory results are likely to persist. As a related

that wants to identify a broader talent pool from among thousands of applicants while staying within employment law boundaries. If your company does not already use algorithms to sort applicants, it will have to do so just to handle the volume of applications.[16] And yet, using algorithms for employment decisions opens the door to another set of criticisms and possible lawsuits.[17] These issues are not likely to go away.

The ongoing call to embrace and pursue DEI raises legal and implementation issues. As a legal matter, companies and other institutions are pursuing diversity goals and/or addressing affirmative action plans; but the two are not the same, and the difference matters.[18] Unlike affirmative action plans, the legal status of diversity plans such as the ones announced by major companies is unclear.[19] As the Equal Employment Opportunity Commission ("EEOC") explains in its Compliance Manual, diversity can be understood as "a business management concept under which employers voluntarily promote an inclusive workplace."[20] Companies have pursued diversity to attract talent and gain "a competitive advantage."[21] In contrast, affirmative action "means those actions appropriate to overcome the effects of past or present practices, policies, or other barriers to equal

example, the Supreme Court is set to hear challenges to college admissions practices at Harvard and the University of North Carolina, where again it considers questions about what is allowed to address diversity and yet not engage in prohibited consideration of race. *See, e.g.*, Nina Totenberg, *The Supreme Court Adds Affirmative Action to Its Potential Hit List*, NPR, https://www.npr.org/2022/01/24/1003049852/supreme-court-adds-affirmative-action-to-its-potential-hit-list (last updated Jan. 24, 2022, 5:39 PM ET) [https://perma.cc/DD8N-JYTA] ("The Supreme Court said Monday it will revisit the question of affirmative action in higher education, deciding to hear cases challenging the use of race as one factor in admissions at Harvard University and the University of North Carolina.").

[16] *See, e.g.*, Sarah K. White & Terena Bell, *Applicant Tracking System: The Secret to Beating a Resume-Filtering ATS*, CIO (Oct. 21, 2021, 2:00 AM PDT), https://www.cio.com/article/2398753/applicant-tracking-system.html [https://perma.cc/D6P2-VEXR] (noting that "75% of recruiters use some type of recruiting or applicant tracking system").

[17] *See* Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 806 n.8 (2020) ("The basic problem of unintentional algorithmic discrimination is by now well-recognized.").

[18] Cynthia L. Estlund, *Putting* Grutter *to Work: Diversity, Integration, and Affirmative Action in the Workplace*, 26 BERKELEY J. EMP. & LAB. L. 1, 4 (2005) (noting the shift to "workforce diversity programs" challenges the classic rationales supporting affirmative action).

[19] *See id.*

[20] Title VII of the Civil Rights Act of 1964, 29 C.F.R. §§ 1600, 1607, 1608 (2022).

[21] *Id.*; *accord* Ajunwa, *supra* note 3, at 1681 n.41.

employment opportunity."[22] Such steps may occur because of a court order, negotiated settlement, or government regulation.[23] Not all affirmative action is mandated. When employers can show a distinct problem such as "a manifest imbalance in a traditionally segregated job category,"[24] employers may use a voluntary affirmative action plan to fix the problem. There is a conceptual and practical link between diversity goals and affirmative action. A company may pursue diversity "for competitive reasons rather than in response to discrimination" and "such initiatives may also help to avoid discrimination."[25] Despite the thin but evolving legal support for DEI plans,[26] the billions of dollars spent on DEI shows that employers are pursing both DEI and affirmative action plans. Methods to support both options are needed.

As another motivation, companies may want to see whether they are missing hiring and talent opportunities.[27] Companies can be stuck in an equilibrium because they rely on, or exploit "old certainties", rather than explore "new possibilities."[28] This exploration/exploitation trade-off began in organizational business literature but has become a significant part of how machine learning ("ML") theory and practice thinks about understanding information.[29] As a matter of best organizational and ML practices, companies need ways explore new candidate pools. Yet, using ML to explore new ways to identify talent raises concerns about algorithmic bias. In December 2020, ten Senators wrote to the EEOC asking it to exercise authority over "hiring technologies" with explicit questions about the use of technology such as machine learning for employment practices, how to review employers' actions based on such

22 EEOC Guidelines on Affirmative Action, 29 C.F.R. § 1608.1(c) (2022).

23 *See* Title VII of the Civil Rights Act of 1964, 29 C.F.R. §§ 1600, 1607, 1608 (2022).

24 EEOC Guidelines on Affirmative Action, 29 C.F.R. § 1608.1(c) (2022).

25 *Id.*

26 *See* Estlund, *supra* note 18, at 14 (noting unwillingness of firms to be a "test case for the legality of diversity-based practices").

27 *Cf.* Susan Dominus, *Tech Companies Face a Fresh Crisis: Hiring*, N.Y. TIMES MAG. (Feb. 16, 2022), https://www.nytimes.com/2022/02/16/magazine/tech-company-recruiters.html [https://perma.cc/8TRL-JKLC] (noting that the tech industry was in need of workers at the time of publication: "tech workers will be recalled as one of the great, pressing shortages of this pandemic"). *See generally* Deven R. Desai, *Exploration and Exploitation: An Essay on (Machine) Learning, Algorithms, and Information Provision*, 47 LOY. U. CHI. L.J. 541 (2015) (explaining businesses operate under the exploration-exploitation tradeoff and exploration is the phase where a business sees whether it is missing opportunities).

28 *See* James G. March, *Exploration and Exploitation in Organizational Learning*, 2 ORG. SCI. 71, 71 (1991).

29 *See* Desai, *supra* note 27, at 568-69.

technology, and what can be done to ensure such technologies do not "deepen systemic patterns of discrimination."[30] The private sector has also raised concerns. For example, The Data & Trust Alliance launched in December 2021, is backed by twenty-two major institutions, and seeks "to detect and combat algorithmic bias"[31] in the hiring process.

Regardless of the motivation behind a company plan, there is a steady drumbeat for algorithmic transparency, especially in employment and college admissions contexts.[32] An entity may have to reveal the process at some point, including in litigation, and need to show that the process is sound from both a mathematical and a legal view.[33] Litigation risks, both real and perceived, could push any company to avoid steps to address diversity. Although some legal scholars argue that certain legal doctrines shield discriminatory outcomes created by using algorithmic processing,[34] debates about what actions are and are not allowed to address diversity persist — especially when using an algorithmic approach.[35]

The situation seems hopeless. One the one hand, the desires and demands for diversity, equity, and inclusion or to pursue affirmative action face legal conundrums about what actions are allowed and what actions

---

[30] Letter from Michael F. Bennet, Cory A. Booker, Sherrod Brown, Elizabeth Warren, Catherine Cortez Masto, Christopher A. Coons, Ron Wyden, Tina Smith, Chris Van Hollen & Jeffrey A. Merkley, U.S. Sens., to Hon. Janet Dhillon, Chair, Equal Emp. Opportunity Comm'n (Dec. 8, 2020), https://www.bennet.senate.gov/public/_cache/files/0/a/0a439d4b-e373-4451-84ed-ba333ce6d1dd/672D2E4304D63A04CC3465C3C8BF1D21.letter-to-chair-dhillon.pdf [https://perma.cc/Z8V9-MXUU].

[31] Steve Lohr, *Group Backed by Top Companies Moves to Combat A.I. Bias in Hiring*, N.Y. TIMES (Dec. 8, 2021), https://www.nytimes.com/2021/12/08/technology/data-trust-alliance-ai-hiring-bias.html [https://perma.cc/SXX9-ZGJG].

[32] *See, e.g.*, Ajunwa, *supra* note 3, at 1680-81 (discussing calls for new legal frameworks to address new technologies in the workplace); Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1 (2017) (examining calls for transparency of software-based decision making).

[33] *Cf.* Bent, *supra* note 17, at 849-50 (noting the need to survive strict or intermediate scrutiny depending on whether race or gender are at issue).

[34] *See* Barocas & Selbst, *supra* note 3, at 709 ("[T]here is good reason to believe that any or all of the data mining models predicated on legitimately job-related traits pass muster under the business necessity defense."); *see also* James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 173-74 (2017) (arguing for a change to business necessity doctrine in light of algorithmic processing in employment practices).

[35] *See, e.g.*, Ajunwa, *supra* note 3 (describing difficulties in the law's ability to redress issues with use of algorithms); Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519 (2018) (discussing if algorithms are the best approach for advancing equality in the workplace); *cf.* Bent, *supra* note 17 (debating the legality of use of algorithms to increase diversity in the workplace).

will result in legal liability. On the other hand, using data and algorithms raises questions about whether one can control for bias and build fair systems to screen and assess candidates. Legal scholarship tends to critique technical solutions as flawed and corrupt while computer science scholarship defers as to whether proposed technical solutions are allowed under the law.[36] Simply put, when entities wish to be proactive regarding diversity, potential discrimination, or to explore whether they have missed opportunities in hiring talent,[37] they will need a path that passes muster against a range of challenges.

This Article bridges the legal and computer science scholarship and shows that employers can use algorithms to pursue DEI or address affirmative action needs and remain within legal rules.[38] We offer legal analysis and computer science techniques to enable companies to improve equal opportunity and employment practices without crossing into arguably illegal discriminatory practices. The Article uses the resume-screening stage of employment to exemplify methods and analysis; but the ideas discussed here are general, and key takeaways can be applied to several stages in the hiring pipeline.

The Article adds to the literature in at least two ways. First, the Article extends the legal literature on whether data analytics and related algorithmic approaches to employment practices are legal.[39] We provide additional legal support for such approaches by showing that the law allows room for employers to address uncertainty in assessing candidates. We also show that the practice of banding candidates as part of evaluating and ranking them is legal precedent that aligns with recent, more

[36] *See* Bent, *supra* note 17, at 806-07, 806 n.8 ("The basic problem of unintentional algorithmic discrimination is by now well-recognized.").

[37] *See* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 118 (2018).

[38] During the writing of this paper, Professor Jason R. Bent reached a similar conclusion regarding the concept of algorithmic affirmative action. *See generally* Bent, *supra* note 17. We broaden this point to address diversity hiring plans as well as affirmative action plans and by offering computer science insights and methods, and we thank Professor Bent for his feedback and encouragement for this work.

[39] *See, e.g.*, Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1039 (2017) (offering general points about algorithms creating "intentional invidious discrimination" and "replicating real world inequalities"); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 13-16 (2014) (criticizing credit-scoring algorithms as entrenching discrimination "in hidden ways"); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 189-90 (2017) [hereinafter *Auditing Algorithms*] (presenting the idea of "classification bias" in algorithms). For an extensive list, see Bent, *supra* note 17, at 806 n.8.

sophisticated computer science and operations research methods available to assess uncertainty in candidates and then select or (partially) sort them.

Second, the Article moves from whether such approaches are allowed to how to pursue such approaches. The Article offers a roadmap for justifying DEI or affirmative action programs. The roadmap shows how to use data-driven and algorithmic techniques to support such action, *and* offers a new method to assess uncertainty in candidates using a mathematical object from 1940s called the partially ordered set ("poset").[40] Using posets offers mathematically robust ways to address issues such as socio-economic differences in test takers (e.g., SAT test takers) and intersectionality issues (e.g., differences among women, minority women, gender orientation) that further complicate how to assess candidates. By extending the legal analysis in support of data-driven and algorithmic approaches to employment and showing how to use those techniques within legal rules, we offer an actionable path towards algorithmic DEI and affirmative action.

Part I sets out the legal rules and data analytics that support DEI and affirmative action practices. Given that affirmative action requirements are developed and specific for each step of building such a plan while the law around diversity only actions are less clear,[41] we use affirmative action law as a guide. Insofar as a DEI plan is likely to be challenged, we take the position that if the plan meets affirmative action standards, it has a good likelihood of being allowed. As a matter of affirmative action, the law supports removing "built-in headwinds" to hiring and promotion practices for minority groups.[42] As such the first step for fairer employment practices is diagnostic; one must show there is something to fix. Part I shows how to use data analytics to make the case that action is warranted.

Part II takes on what is allowed to address such headwinds. It unravels the tensions between actions that may be seen as disparate treatment or disparate impact outcomes and revisits *Ricci v. DeStefano* to show *Ricci* fully supports steps to design fairer assessment methods and engage in bias mitigation. This point leads to the next question. What sorts of bias mitigation is allowed?

Part III turns to that question by way of computer science and sets out current approaches to bias mitigation. It investigates such approaches and their limits. This Part then introduces and explains a new method, the

[40] GARRETT BIRKHOFF, LATTICE THEORY 1, 5 (1940).

[41] *See* Estlund, *supra* note 18, at 21.

[42] *See* Ricci v. DeStefano, 557 U.S. 557, 622 (2009) (Ginsburg, J., dissenting).

partially ordered set or poset method, to address uncertainty and fairness in assessing candidates. Part III also offers two examples of the approach in action.

Part IV returns to the law and current attempts to protect algorithmic hiring. It shows that these efforts, although laudable, fall short of desired outcomes, and as the recent news about the Rooney Rule in the NFL shows, can end up being quotas for interviewing that devolve into tokenism rather than true consideration.[43] Part IV then shows how the poset approach enables better comparisons of candidates (thus moving away from tokenism) and enables an algorithmic diversity, equity, and inclusion. The Article then concludes with observations about the implications of this work.

I. USING DATA ANALYSIS TO IDENTIFY AND SUPPORT DEI AND AFFIRMATIVE ACTION PRACTICES

An employer concerned that its workforce under-represents women and minorities faces a seemingly simple question: May they do anything to change their current hiring practices? The answer appears to be yes, but the exact nature of what steps are allowed is where companies run into problems. The first question is what supports the claim that corrective action is needed? Both the law and machine learning offer an answer.

The purpose behind Title VII is "to achieve equality of employment opportunities," and Congress "directed the thrust of the Act to the consequences of employment practices, not simply the motivation."[44] That means "unnecessary barriers to employment" must fall, even if "neutral on their face" and "neutral in terms of intent."[45] Federal courts have disallowed a host of hiring and promotion practices that "operate[d] as 'built in headwinds' for minority groups."[46] In addition, the Supreme Court has upheld the legality of employment plans to address discrimination without reference to its past practices or evidence of a possible violation of the law.[47]

To act, an employer "need[s] to point only to a 'conspicuous . . . imbalance in traditionally segregated job categories.'"[48] Logically, this

43 *See infra* Section IV.A.3.
44 Griggs v. Duke Power Co., 401 U.S. 424, 429-32 (1971).
45 *Id.* at 430-31.
46 *Ricci*, 557 U.S. at 622.
47 Johnson v. Transp. Agency, 480 U.S. 616, 630 (1987).
48 *Id.*

requirement allows initial, proactive analysis identifying the imbalance problems. This Part sets how to use data science and analytics to identify imbalances in the hiring pipeline that a company seeks to address,[49] and how to address to possible sources of an imbalance so that an employer can see where corrective action is needed.

A. Demonstrating Inequity — Diagnosing Imbalance

Contrary to scholarship that tells tales of data leading to negative results or necessarily having a disparate impact,[50] an employer can and should use data analytics to examine its hiring and employment practices. First, an employer can audit its current workforce and get fine-grained information about who works at the company and at what levels. Such an approach allows the company to look beyond simple questions such as "Does it have an equal number of men and women or minorities in the workforce?" Instead, the company can see the gender and minority makeup at different levels of employment such as upper management, upper-middle management, middle management, administration, hourly workers, contractors, and so on. Depending on goals or concerns, an audit can help see whether practices raise intersectionality issues such as whether women are hired but minority women are not. Visualizing the data with pie-charts or heat maps will provide clear, vivid ways to see the current situation. Second, after such a study, the company is set up to see where potential sources of issues arise. It may find that women and minorities rarely move beyond middle management or are rarely interviewed for promotion. Or it might detect that its screening tools are skewing the intake process. Diagnostics must be used to understand the status quo, but more is needed. The data must also show there is an imbalance.

Best practices in data-science show how to identify imbalances. Public service offers a good example. In one of the earliest examples of data analysis and visualization, Dr. John Snow detected the source of cholera

[49] Kimberly A. Houser, *Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making*, 22 STAN. TECH. L. REV. 290, 324 (2019) (arguing for "responsible use of AI" to address employment bias); Kim, *Auditing Algorithms*, *supra* note 39, at 197 (noting possibility of modifying algorithms prospectively to address bias); Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67, 125-29 (2017-18) (examining whether caselaw allows changes in algorithms to address disparate impact).

[50] *See* Barocas & Selbst, *supra* note 3, at 685. *See generally* O'NEIL, *supra* note 3, at 11-13 (highlighting that "ill-conceived mathematical models" and "rogue algorithms" control college admissions, ending, sentencing, and employment via "secret models wielding arbitrary punishments").

and thus enabled actions to defeat the deadly disease.[51] London had experienced several outbreaks with one in 1849 causing about 53,000 deaths.[52] The 1854 outbreak had a high death rate with 600 deaths in one week that September.[53] There were two competing theories about the cause. The dominant theory, the miasma theory, held that "cholera was caused by airborne transmission of poisonous vapors from foul smells due to poor sanitation."[54] Given the nature of London sewage systems and the way they fouled the Thames River, that theory is not surprising.[55] The other theory, the Germ Theory, was "an unproven minority opinion in medical circles" to which Snow subscribed.[56]

Snow used data to identify the cause of the outbreak and persuade authorities about how to address the problem. During the August to September outbreak in 1854, Snow documented where cholera deaths occurred over a seven-week period.[57]

TABLE IX.

	Number of houses.	Deaths from Cholera.	Deaths in each 10,000 houses.
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company . . .	26,107	98	37
Rest of London	256,423	1,422	59

He charted the water supply, the houses, the deaths from cholera, and the deaths per 10,000 houses.[58] Snow dug deeper into his data and showed that "brewery workers and poorhouse residents in the area, both of whom relied on local wells, escaped the epidemic."[59] In addition, he created two

51 *See generally* THEODORE H. TULCHINSKY, *John Snow, Cholera, the Broad Street Pump; Waterborne Diseases Then and Now*, *in* CASE STUDIES IN PUBLIC HEALTH 77 (2018) (detailing Snow's work to identify the source of a cholera outbreak in England in the 1850s).

52 *See id.* at 80.

53 *Id.*

54 *Id.*

55 *Id.*

56 *Id.*

57 *Id.* at 80-82.

58 JOHN SNOW, ON THE MODE OF COMMUNICATION OF CHOLERA 55-98 (2d ed. 1855).

59 TULCHINSKY, *supra* note 51, at 81.

maps, what today would be called visualizations, to show the stark differences in cholera deaths.[60] The numbers and maps led to his "observation that the cases either lived close to or were using the Broad Street pump for drinking water."[61] Snow then was able to assert the problem stemmed from the water supply. Those who had access to uncontaminated water did not get cholera, but "users of the Broad Street pump became infected."[62] Using data and visualizations, Snow convinced the authorities to get rid of the Broad Street pump, accelerating the end of the epidemic.[63]

In a less severe city context, one city's effort to find and fix potholes shows how good data practices can lead to good outcomes. The City of Boston deployed an app to help detect potholes and fill them fast. The app had problems including design issues — a requirement to launch the app at the start of the trip and close it at the end, plus the app could not run in the background so other apps like Google Maps would not work — that deterred people from using the app.[64] A criticism is that the app would select for the wealthy as they would be the ones with smart phones. The reality was that few people used the app other than city workers who were required to use it. The limited use might have led to underrepresentation, *and yet* Boston appears to have avoided that outcome.

Unfortunately, parts of academia have used Boston and its pothole app experience to give the impression of harm rather than embracing the results of good data practices. For example, Barocas and Selbst's paper *Big Data's Disparate Impact* discusses the app in the section, "How Data Mining Discriminates," to support the idea the big data discriminates.[65] They argue, "systematic differences in smartphone ownership will very likely result in the underreporting of road problems in the poorer communities where protected groups disproportionately congregate. *If* the city were to rely on this data to determine where it should direct its

[60] *See* Fahema Begum, *Mapping Disease: John Snow and Cholera*, ROYAL COLL. OF SURGEONS OF ENG. (Dec. 9, 2016), https://www.rcseng.ac.uk/library-and-publications/library/blog/mapping-disease-john-snow-and-cholera/ [https://perma.cc/9SD5-38G8].

[61] TULCHINSKY, *supra* note 51, at 81.

[62] *Id.*

[63] *See id.*

[64] *See* CEA, *Street Bump: Crowdsourcing Better Streets, but Many Roadblocks Remain*, HARV. DIGIT. INNOVATION & TRANSFORMATION (Oct. 30, 2015), https://d3.harvard.edu/platform-digit/submission/street-bump-crowdsourcing-better-streets-but-many-roadblocks-remain/ [https://perma.cc/6SBX-XV7Q].

[65] Barocas & Selbst, *supra* note 3, at 685.

resources, it would only further underserve these communities."[66] In discussing the app and "The Difficulty for Reforms" Barocas and Selbst claim:

> In many cases, however, an analyst can only determine the extent of—and correct for—unintentional discrimination that results from reporting, sampling, and selection biases if the analyst has access to information that somehow reveals misrepresentations of protected classes in the dataset. Often, there may be no practical alternative method for collecting information that would reveal the existence of a bias.[67]

Three problems flow from this view of data.

First, data are inanimate and lack agency. Data cannot discriminate. Data cannot do anything. Second, whether the use of data leads to undesired outcomes rests on whether the data user would accept the outcome of the data-driven project or ask whether the outcome made sense. If Boston had rested on the idea that the app and data was magically perfect the issues might have arisen, but Boston used the data to address gaps in the project. Boston avoided what Professor Ajunwa has called "a false binary . . . . willfully forget[ing] that the human hand remains present in all automated decision-making."[68]

Third, the idea that little can be done to address errors in data is at least overstated.[69] As discussed below, statistics and computer science have a rich literature focused on addressing errors and potential biases in data. In addition, one solution is simpler and focuses on the people using data and software. As Barocas and Selbst finally acknowledge but dismiss as uncommon, Boston *was* conscientious, and its Office of New Urban Mechanics worked to identify and address gaps.[70] That partnership

---

[66] *Id.* (emphasis added).

[67] *Id.* at 718.

[68] Ajunwa, *supra* note 3, at 1681.

[69] Barocas & Selbst acknowledge a technique — pre-processing of data — and retroactive data correction are possible ways to address data mining issues. *See* Barocas & Selbst, *supra* note 3, at 719. But the overall thrust of the claim is that so many steps in data mining are difficult that legal reforms will falter because "policies that compel institutions to correct tainted datasets or biased samples will make impossible demands of analysts." *Id.* at 722. And they conclude their analysis by implying we are stuck with disparate impact even when companies take steps to address the problem. *Id.* ("[E]ven when companies voluntarily adopt such strategies, these internal difficulties will likely allow a disparate impact to persist.").

[70] *See id.* at 718; *accord* Kate Crawford, *Think Again: Big Data*, FOREIGN POL'Y (May 10, 2013, 12:40 AM), https://foreignpolicy.com/2013/05/10/think-again-big-data/

allowed experts to work with the city to identify and address problems. In the specific case, the app first detected manhole covers as potholes and that false report had to be accounted for. In addition, a data process that connects to geographic patterns lends itself to visualization to understand the results of the data analytic practice.[71] A map with the data overlaid should reveal oddities.

Automated decision-making can be complicated; it may not be apparent from observing the decision-making process whether decisions will be fair. Outcomes, however, can point to potential problems. By analogy, a layperson may not glean any information by looking at a car's engine; however, an unusual sound while driving is a signal to have the car assessed.

If only part of Boston seems to have potholes, even though the entire city is covered with snow, ice, salt, and sand in the winter, those in charge should be startled, or least curious, about the result. Digging deeper might reveal that wealthy areas *seem* to be the only areas with potholes. One might be lazy and rationalize that rich people drive over-sized vehicles and tend to ignore rules about snow tires, and so the rich, possibly white, areas will have more potholes to fix. But that conclusion misses the better ways to use data — questioning and testing odd results rather than deferring to them.

An analyst needs information that reveals possible problems with results, but the idea that there is no practical alternative to "reveal the existence of bias" misses the way data science can operate. That position — assuming data practices are a one-off use and unquestionable is ironic for two reasons. First, the source of the position comes from the critique that wants people not to defer to data and yet holds that such deference is almost inevitable. Second, the position cedes the agency and responsibility of those who use data and software, a position that underappreciates best practices in data science, the approaches to bias mitigation that exist, and so seems to take the position that horrible outcomes are inevitable. In contrast, as Professor Bornstein has summed up, "Whether an algorithm could result in exacerbating or, alternatively, reducing protected class biases will depend on both how it is created and how it is used."[72]

[https://perma.cc/27GU-MFHC] ("Boston's Office of New Urban Mechanics has made concerted efforts to address these potential data gaps.").

71 Barocas & Selbst, *supra* note 3, at 717-18 (acknowledging that Boston used visualization which aided in detecting a problem in their approach to potholes).

72 Bornstein, *supra* note 35, at 553; *cf.* WILLIAM W. LOWRANCE, MODERN SCIENCE AND HUMAN VALUES 37 (1986) ("It makes no sense, though, to attribute to technology a mind of its own. . . . 'like all bad workmen we blame our tools.'" (quoting Peter Medawar)).

Put differently, a "take it for granted that the tool is working well approach" should not be accepted as valid. Auditing and interrogating outcomes so that anomalies are found should prompt a check of what is going on in the areas that are not detecting potholes or when women are not being interviewed.[73] As one of us tells their class, "Say that out loud. There are no potholes in major parts of Boston. There are no qualified women to interview for this position. Do these statements make sense?" The proper general concern is that someone using such methods could fail to use them well, not that they used data and related analytical techniques in the first place.[74]

In the hiring context, a data-analytic approach that audits and assesses the status quo allows a company to see the full range of the effects of its employment practices. In another example used to scare people about data, legal scholars point to Amazon's development of a resume-screening algorithm that penalized resumes which included the word "women's" due to data of past hiring trends in the company.[75] The algorithm thus would under-value those who attended all-women colleges and reward vocabulary typically used by men. What scholars missed is that Amazon developed the tool, reviewed the outcomes, detected the odd outcomes, and so *did not* use it.[76] Although one can be concerned about the potential harm from Amazon's initial system, an equally important lesson is that the company used good practices and thus detected imbalances.

Auditing hiring processes — while important in all contexts — is especially important in the context of data-driven systems. Apart from

[73] *Cf.* Sam Corbett-Davies, Emma Pierson, Avi Feller & Sharad Goel, *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016, 5:00 AM EDT), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ [https://perma.cc/R4GN-2M79] ("[S]ince classification errors here disproportionately affect [B]lack defendants, we have an obligation to explore alternative policies.").

[74] *See, e.g.*, Crawford, *supra* note 70 ("[L]ess conscientious public officials may miss ['data gaps'] and end up misallocating resources in ways that further entrench existing social inequities.").

[75] *See* Bornstein, *supra* note 35, at 521 (discussing the automated tool developed by Amazon to rank candidates to automate hiring); Ajunwa, *supra* note 3, at 1673-74; Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (last updated Oct. 10, 2018, 4:04 PM) [https://perma.cc/KAS8-UNTA].

[76] Dastin, *supra* note 75; *accord* Bent, *supra* note 17, at 806 ("Amazon gave up on that project . . . .").

systemic issues such as historic discrimination, gender and racial norms, demographically imbalanced datasets, and so on, we must simply contend with the fact that data are messy. In a working paper, Salem and Gupta conducted a case study in which they randomly sampled a training set from a larger dataset (it is a standard practice in machine learning to split the data by the 80-20 rule into a training and testing split). By chance, in their example, the average score of women in the training set was lower than the average score of women in the larger dataset. A training set is never a perfect representation of future applicant pools. Even if the training set were perfectly demographically proportional, it might be the case that the representatives of some demographic group in the training set had sub-par outcomes due to random chance. This group might then be treated poorly in future rounds of hiring due to this trend in the data. Moreover, their toy machine learning model picked up a skew benefitting males in the data by assigning them a slightly higher score. A selection algorithm based on such an ML model would then select men more frequently than women. In this work, Salem and Gupta further show that a closer look into the variables of their (linear regression) model included "gender" which was assigned -16.95 points. Female candidates had this variable set to 1, and therefore their predicted score was -16.95 points less than male candidates who had all other variables equal. Of course, one could ask that "gender" not be included in the prediction, but this does not fix the problem of an ML model picking up an undesirable trend in the data. A similar skew could have been learned by a neural network, or a support vector machine, even without the consideration of "gender." As discussed later, Salem and Gupta used partially ordered sets or "a poset approach" to account for such skews, as best as possible, by allowing for an uncertainty around the predicted score.[77]

Using good data science methods to audit should allow an entity to document "conspicuous . . . imbalance in traditionally segregated job categories."[78] Rather than resting easy and deferring to the data and software in place, the next step is to identify aspects of its employment practices that create "unnecessary barriers to employment,"[79] or "operate as 'built in headwinds' for minority groups,"[80] and so support the case that there is something to fix.

[77] *See infra* note 166 and accompanying text.

[78] Johnson v. Transp. Agency, 480 U.S. 616, 630 (1987).

[79] Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

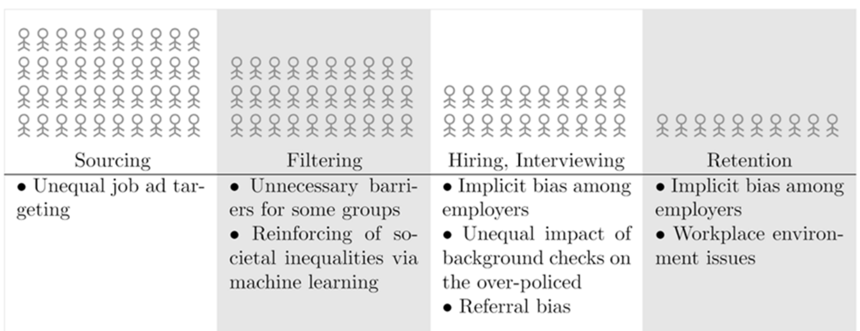
[80] Ricci v. DeStefano, 557 U.S. 557, 622 (2009) (Ginsburg, J., dissenting).

B. Identifying Built-in Headwinds

1. The General Problem: Sorting for Talent

An employer seeking to improve diversity, implement an affirmative action plan, or explore new sources of talent will need to use advertising to announce positions and hope its ads reach viable candidates. The employer may face thousands of applicants for the position or perhaps a few hundred for more senior positions. In either case, the employer needs a way to sort resumes to see who to interview. Interviewing and hiring will require ways to assess and often rank candidates. After a hire is made, an employer will face a similar set of challenges as it sorts retention and promotion practices, especially if the company is a large one such as Amazon or Walmart. The problem is that there are a number of junctures in the hiring pipeline at which bias can affect decisions, as depicted in Figure 1.[81]

Figure 1: Some ethical concerns in various stages of the employment pipeline.



Job advertisements on various platforms can be targeted at specific audiences.[82] Application rates can differ across groups due to presumed

[81] These stages are well-known and at least one other study of the use of technology and hiring uses the same stages but under different names. *See* MIRANDA BOGEN & AARON RIEKE, HELP WANTED: AN EXAMINATION OF HIRING ALGORITHMS, EQUITY, AND BIAS 13 (2018).

[82] Pauline T. Kim, *Manipulating Opportunity*, 106 VA. L. REV. 867, 870-71 (2020) (noting online intermediaries' ability for precise targeting of ads); Julia Angwin, Noam Scheiber & Ariana Tobin, *Dozens of Companies Are Using Facebook to Exclude Older Workers from Job Ads*, PROPUBLICA (Dec. 20, 2017, 5:45 PM EST), https://www.propublica.org/article/facebook-ads-age-discrimination-targeting [https://perma.cc/V9MF-KYVR].

employer bias.[83] Data-driven tools for evaluating resumes can be biased due to inequalities in hiring data,[84] imbalance in data,[85] or differences in false positive/negative error rates in prediction algorithms leading to bias as a *downstream effect*.[86] Referral hiring can lead to favoritism.[87] Customer evaluations of freelancers can adversely impact certain groups.[88] Final hiring decisions can be influenced by human biases of the hiring committee.[89] After going through the hiring pipeline, candidates also see a significant difference in salaries offered,[90] and retention rates can differ dependent on the work environment.[91] Indeed, societal biases are

[83] Tara Sophia Mohr, *Why Women Don't Apply for Jobs Unless They're 100% Qualified*, HARV. BUS. REV. (Aug. 25, 2014), https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified [https://perma.cc/72RV-Z5PZ].

[84] Rachel Goodman, *Why Amazon's Automated Hiring Tool Discriminated Against Women*, ACLU (Oct. 12, 2018), https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against [https://perma.cc/TC8X-8T8L].

[85] *See* Seyma Yucer, Samet Akçay, Noura Al-Moubayed & Toby P. Breckon, *Exploring Racial Bias Within Face Recognition via Per-Subject Adversarially-Enabled Data Augmentation*, 2020 IEEE/CVF CONF. ON COMPUT. VISION & PATTERN RECOGNITION WORKSHOPS 83, 83.

[86] *See* Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain & Lucy Vasserman, *Measuring and Mitigating Unintended Bias in Text Classification*, 2018 AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y 67, 71 (noting that imbalance in training data can result in biased decisions for text classification tasks. Generally, textual machine learning settings involve several steps, from pre-processing data, to training a general model, to fine-tuning the general model for a specific application. Biases at one of these stages can propagate to later stages).

[87] Steven D. Schlachter & Jenna R. Pieper, *Employee Referral Hiring in Organizations: An Integrative Conceptual Review, Model, and Agenda for Future Research*, 104 J. APPLIED PSYCH. 1325, 1341 (2019).

[88] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier & Christo Wilson, *Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr*, 2017 ACM CONF. ON COMPUT. SUPPORTED COOP. WORK & SOC. COMPUTING 1914, 1914-15.

[89] Ashley B. Batastini, Angelea D. Bolaños, Robert D. Morgan & Sean M. Mitchell, *Bias in Hiring Applicants with Mental Illness and Criminal Justice Involvement: A Follow-up Study with Employers*, 44 CRIM. JUST. & BEHAV. 777, 784-86 (2017).

[90] Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham & Jo Handelsman, *Science Faculty's Subtle Gender Biases Favor Male Students*, 109 PROC. NAT'L ACAD. SCIS. 16474, 16475 (2012).

[91] Vedant Das Swain, Koustuv Saha, Manikanta D. Reddy, Hemang Rajvanshy, Gregory D. Abowd & Munmun De Choudhury, *Modeling Organizational Culture with Workplace Experiences Shared on Glassdoor*, 2020 CHI CONF. ON HUM. FACTORS IN COMPUT. SYS. 1, 9.

pervasive and can affect decisions made by experts.[92] At each of these stages, some type of algorithm — a system to manage, sort, and assess — is often used.

Although the recent attention to algorithms[93] and employment may make it seem like a new approach, the practice can be traced back at least 40 years.[94] When automated systems are used at any stage, missed opportunity (false negatives) with respect to minority candidates is often shrugged off as an artifact of the prediction model, necessary for overall accuracy.[95] These models often train on historic data, which can depict imbalanced selection rates across different groups of candidates, and these trends can be learned by automated methods.[96] History can dictate future actions. In short, existing pipeline practices can reiterate and increase disparity in opportunity and outcomes.

2. Resume Screening: A Place for Intervention

Although the hiring pipeline can be improved in many places, we examine the resume-screening stage as particularly ripe for improvement[97] and focus the Article on this stage for several reasons. First, it is a good

[92] CRAIG HANKS, TECHNOLOGY AND VALUES: ESSENTIAL READINGS 41 (2009).

[93] The term algorithm has come to inspire fear and assumptions that what is at issue are opaque magical things, but they are not. *See, e.g.*, Desai & Kroll, *supra* note 32, at 4 (dispelling myths about the nature of algorithms and that problems associated with algorithmic transparency are aggravated by a lack of technical understanding). Everything from a recipe to a deep learning system are algorithms, and it is best in our context to think of the issues as relating to software. *See id.* at 23-30.

[94] Oscar Schwartz, *Untold History of AI: Algorithmic Bias Was Born in the 1980s*, IEEE SPECTRUM (Apr. 15, 2019), https://spectrum.ieee.org/tech-talk/tech-history/dawn-of-electronics/untold-history-of-ai-the-birth-of-machine-bias [https://perma.cc/8H4L-VACH]; James Hu, *Over 98% of Fortune 500 Companies Use Applicant Tracking Systems (ATS)*, JOBSCAN (June 20, 2018), https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/ [https://perma.cc/Y65G-6X5Z].

[95] MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN 75 (2020).

[96] *See* Barocas & Selbst, *supra* note 3, at 674 ("Approached without care, data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society."); Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCIENCE 183, 185 (2017).

[97] *Cf.* Naomi Nix, *Removing Résumés from Hiring Process Can Improve Diversity*, BLOOMBERG (Feb. 21, 2022, 3:00 AM PST), https://www.bloomberg.com/news/articles/2022-02-21/removing-work-r-sum-s-from-job-hiring-process-can-improve-candidate-diversity [https://perma.cc/U3C6-K6CJ] (discussing companies using screening methods other than resumes to identify talent).

lens through which to investigate the concerns around using algorithms and data in the employment context. There may be hundreds if not thousands of applications for one or a handful of open positions. Manual analysis of a high volume of resumes to select candidates who should advance to the next round of the hiring process is not viable. Using algorithms to screen and sort thousands of resumes, however, makes the problem manageable.

This approach has several appealing advantages: speed, cost-effectiveness, potential objectivity, and uniformity in process. These properties may seem desirable from an ethical and fairness perspective. Consistency in decisions is often a good thing, and a lack of human involvement would seem to minimize the role of implicit bias in hiring decisions.[98] But the very nature of such processes poses another problem. As with other stages in the hiring process, because seemingly objective methods interact with real-world data, automated decisions can *reflect* and therefore, *reinforce* societal inequalities.[99] Even when there is no intent to discriminate, and the decision system uses the same data and applies the same rule to all, there may be a disproportionate effect on a protected class (i.e., groups protected by law from discrimination, such as those defined by sex, race, age, etc.).[100] In short, the problems in resume screening map to the more general ones present when using data-driven decision-making in the employment context.

Second, algorithms are already used for screening applications. This practice creates an advantage, as certain techniques can be used to address — rather than propagate — bias. Adjusting algorithmic techniques may be a more palatable idea and possible in an industry currently using automated processes than using algorithms in a heretofore un-automated process. New algorithmic interventions are, therefore, more likely to be applied in practice.

---

[98] *See, e.g.*, Rema N. Hanna & Leigh L. Linden, *Discrimination in Grading*, 4 AM. ECON. J.: ECON. POL'Y 146, 158 (2012) (reporting study results from a group of teachers and students in India that show the teacher discrimination was present in grading but with a relatively small effect); *cf.* Ajunwa, *supra* note 3, at 1686 (discussing Professor Kate Crawford's argument about viewing data as purely objective, clarifying that biases can be as prevalent in big data as much as they are in individual perceptions).

[99] Benjamin Edelman, Michael Luca & Dan Svirsky, *Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment*, 9 AM. ECON. J.: APPLIED ECON. 1, 7 (2017).

[100] Kristian Lum & William Isaac, *To Predict and Serve?*, 13 SIGNIFICANCE 14, 18 (2016); Solon Barocas, *Data Mining and the Discourse on Discrimination*, *in* DATA ETHICS WORKSHOP, CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 1-4 (2014).

Third, changes at early stages of the hiring pipeline are vital to address later bias. Changes at later stages are only meaningful if they act on a diverse pool of candidates. Without a diverse candidate pool at those stages, efforts to address bias become empty theater, because there will be few to no candidates from underrepresented groups for which the changes would help. As such, we focus specifically on automated resume-screening processes: *how should applicant-screening methods be developed?*

3. Identifying Bias in Resume Screening

We broadly refer to systematic inconsistencies in data that adversely affect certain groups as "bias." For example, a study was conducted in which resumes were sent to science faculty to be evaluated.[101] These resumes were identical but for the name, which was either John or Jennifer, and the resumes of John received a higher average score. Similar studies have been conducted with similar results, such as the racial experiment of Bertrand and Mullainathan.[102] These are quite blatant examples of bias. More subtle examples are discussed below.

Bias is a leaning for which one must account.[103] Attempts to mitigate bias often begin with an understanding of the nature of the bias, or in other words, the inconsistencies in measurement of the ability of candidates. Unfair decisions can stem from many places and identifying the origins of the bias allows for precise interventions. In the hiring process (automated or otherwise),[104] applications will typically be assigned a score, thus allowing comparisons of applicants based on a single number or with respect to a single ranking of candidates.[105] When using an algorithm, this

---

[101] Moss-Racusin et al., *supra* note 90.

[102] Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991, 998 (2004).

[103] LOWRANCE, *supra* note 72, at 7 ("'[B]ias' simply means inclination, and isn't necessarily pejorative."); *id.* at 67 (documenting the way each decision in biological sciences such as which genetic strain of rodent to use in a study, test choices by toxicologists, which tissue samples a pathologists selects, creates a bias).

[104] *See* Bradley v. City of Lynn, 443 F. Supp. 2d 145, 168 (D. Mass. 2006) ("The effect of using examination scores, which disparately impact minorities at all scores above seventy, for rank ordering, is to bunch minorities at the bottom of the eligible list.").

[105] Javier Sánchez-Monedero, Lina Dencik & Lilian Edwards, *What Does It Mean to "Solve" the Problem of Discrimination in Hiring? Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems*, 2020 CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 458, 460, 463; *Applicant Tracking Systems*, JOBSCAN,

evaluation metric can be hard coded into an algorithm or developed dynamically, and in either case, can be unfair. A natural question is whether we can model this bias precisely and account for it within the algorithms to make them justifiably (provably) fairer.

The challenge in modeling biased evaluations is that such bias can take different forms and be observed in different ways. As discussed above, a resume-screening algorithm developed, but not employed by, Amazon, penalized resumes that included the word "women's" because the system drew on resumes of current and past employees who tended to be male.[106] Note that this form of unfairness — while clear from a definitional sense — can be hard to observe in practice, as applicants are never truly identical but for a small number of attributes. Even if Amazon had used its tool, other factors such as grade point average, degree obtained, school attended, and so on might make it seem that gender was not having the effect that was detected.[107] Both examples are clear examples of bias, as toggling a protected attribute results in different treatment. The precise cause may be unclear (such results may come from the effects of implicit bias in past hiring decisions, gender imbalance in training datasets, gender norms pigeonholing people into certain academic and career paths, etc.), but some systemic issues have resulted in evaluations skewed against women.

Many cases of bias in evaluations are, however, more nuanced. Consider using SAT scores to screen candidates — a practice employers such as McKinsey, Bain, Goldman Sachs, and Amazon have been known to use even for candidates with advanced degrees.[108] The problem is that testing results are not ironclad indicators of ability. They are not ground truth. Studies show that even when students are equally able to perform well on a test, if the test is announced to exhibit differences across groups, students in a negatively stereotyped group perform lower than the students in a non-

https://www.jobscan.co/applicant-tracking-systems (last visited Dec. 8, 2020) [https://perma.cc/FRP2-U8XJ].

[106] Dastin, *supra* note 75.

[107] In a similar vein, an empirical study showed that science faculty's assessment of resumes varied dependent on the gender of the student. *See* Moss-Racusin et al., *supra* note 90.

[108] Shaila Dewan, *How Businesses Use Your SATs*, N.Y. TIMES (Mar. 29, 2014), https://www.nytimes.com/2014/03/30/sunday-review/how-businesses-use-your-sats.html [https://perma.cc/JB3L-EA74]; *McKinsey's Online Application FAQs*, MCKINSEY & CO., https://www.mckinsey.com/careers/application-faq (last visited Dec. 22, 2022) [https://perma.cc/K2W5-NQF5]; Alison Griswold, *Why Major Companies Like Amazon Ask Job Candidates for Their SAT Scores*, YAHOO! NEWS BUS. INSIDER (Mar. 4, 2014), https://sg.news.yahoo.com/why-goldman-sachs-bain-mckinsey-170407444.html [https://perma.cc/C829-QNFJ].

stereotyped group.[109] That is, once a group believes they are unlikely to do well on the test, that belief affects actual performance. Another study from 2013 shows that SAT scores are correlated with family income, potentially pointing to issues of access.[110] Inside Higher Education looked at SAT scores in 2015 and found that despite fee waivers and increased efforts to provide support and tutoring to low-income families:

> In each of the three parts of the SAT, the lowest average scores were those with less than $20,000 in family income, and the highest averages were those with more than $200,000 in income, and the gaps are significant. In reading, for example, the average for those with family income below $20,000 is 433, while the average for those with income of above $200,000 is 570.[111]

Thus, despite steps to address economic inequality's effect on testers, compared to 2013, gaps in performance with respect to racial groups not only persisted, but increased. This problem with SAT scores is further evident in a recent study by Faenza, Gupta, and Zhang,[112] which showed a shift by approximately 200 points in SAT scores from schools with different economic need indices. Thus, an employer using SAT scores appears neutral but sets up a pre-selected pool.

As a simple example, consider Figure 2. Suppose a hiring committee wants to select two of the applicants represented in the right plot. If the method to generate the scores has biases or uncertainties depending on gender, and if only the raw evaluations (the centers of the intervals) are used to make these decisions (Figure 2, right), then only the two high-scoring male candidates could be selected, as they are the only applicants meeting the cutoff. The raw score approach fails to account for the uncertainty in the scoring method and so the persons just below the cutoff are unnecessarily and erroneously not in the selection pool.

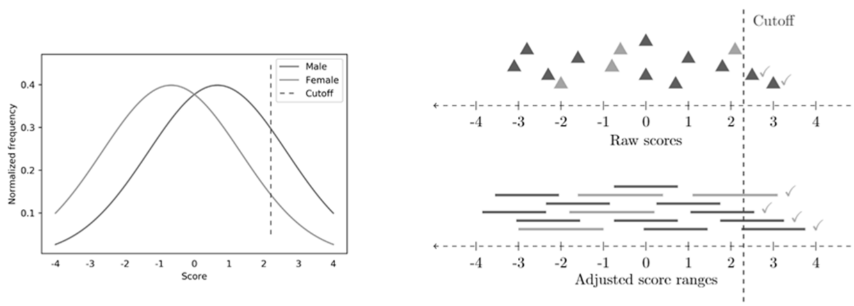
[109] Claude M. Steele & Joshua Aronson, *Stereotype Threat and the Intellectual Test Performance of African Americans*, 69 J. PERSONALITY & SOC. PSYCH. 797, 800 (1995).

[110] Ezekiel J. Dixon-Román, Howard T. Everson & John J. McArdle, *Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance*, 115 TCHRS. COLL. REC. 1, 22 (2013).

[111] Scott Jaschik, *SAT Scores Drop*, INSIDE HIGHER ED (Sept. 3, 2015), https://www.insidehighered.com/news/2015/09/03/sat-scores-drop-and-racial-gaps-remain-large [https://perma.cc/7QVS-8VTN].

[112] Yuri Faenza, Swati Gupta & Xuan Zhang, *Impact of Bias on School Admissions and Targeted Interventions*, arXiv preprint arXiv:2004.10846 (2020).

Figure 2: (left) Example of predicted score distributions by gender (blue: male, orange: female) and (right) potential score ranges for candidates from these distributions which arguably contain their "true" score with high probability.[113]



Extrapolating from Figure 2, a negligible shift in one group's distribution (potentially caused by a bias) can result in a large difference selections rates across genders. When decisions are made at scale and uncertainties are not considered, the disparate effect across groups can become significant.

Identified, strong evidence of bias in current algorithmic sorting in the hiring process, including resume screening, should constitute the sort of "built in headwind[] for minority groups" that the law seeks to eliminate.[114] With sufficient evidence of bias and systemic barriers to equality of employment opportunities, an employer can make a case for using bias-aware algorithms, because it will have mathematical evidence of a clear "unnecessary barrier to employment" even if the system is "neutral on its face."[115] Finding such evidence leads to a new question: what does the law allow an entity to do to address the problems?

II. RE-READING *RICCI* OR WHAT CAN ONE DO TO REDUCE BUILT-IN HEADWINDS?

Voluntary action to comply with the goals of Title VII is not only allowed; it is favored.[116] Nonetheless, in some cases, trying to further the goals of Title VII to address discrimination raises the paradox where one

[113] Jad Salem, Deven Desai & Swati Gupta, *Don't Let* Ricci v. DeStefano *Hold You Back: A Bias-Aware Legal Solution to the Hiring Paradox*, 2022 CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 651, 665 fig.4.

[114] Ricci v. DeStefano, 557 U.S. 557, 622 (2009) (Ginsburg, J., dissenting).

[115] Griggs v. Duke Power, 401 U.S. 424, 430 (1971).

[116] Johnson v. Transp. Agency, 480 U.S. 616, 630 (1987).

approach looks like disparate impact and a corrective action looks like disparate treatment. This Part sets out whether an entity may use protected class information to increase diversity among interviewees.

A. The Disparate Impact, Disparate Treatment Trade-off

If an employer uses an algorithmic tool to evaluate and screen candidates, the employer may face legal challenges depending on the outputs of the tool. Recall that an algorithmic tool can have biases even if the tool was not designed to discriminate. That's what happened with Amazon's unused system. The system used rich data about past success to sort candidates, and luckily, Amazon detected the problem. Now, imagine Amazon had used the system. The program would have yielded undesired results, but the programmers could say that the program was not designed to act that way. This problem tracks the disparate impact doctrine in anti-discrimination law.

Disparate impact addresses when "facially neutral policies or practices have a disproportionate adverse effect or impact on a protected class . . . ."[117] Disparate impact doctrine is thus supposed to address situations where intent is not at hand or cannot be ascertained.[118] Outcomes based on unaware algorithms may fit quite well with disparate impact challenges, because unaware algorithms are facially neutral, the software designers may lack intent to discriminate, and nonetheless the software yields statistically discriminatory results.

The possibility of a disparate impact claim leads to an obvious, yet problematic, approach. An employer may design a more aware algorithm that takes protected class status into account. And yet this approach may run into a disparate treatment challenge. That doctrine prohibits intentionally using race, gender, or other protected class status to make decisions about credit, employment, housing, and other regulated areas of

---

[117] FED. TRADE COMM'N [FTC], BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? 19 (2016) (citing 12 C.F.R. § 1002.6 (2023) (citing Griggs v. Duke Power Co., 401 U.S. 424 (1971), and Ablemarle Paper Co. v. Moody, 422 U.S. 405, 430-31 (1975))); *accord* 42 U.S.C. § 2000e-2(k)(1)(A) (2018).

[118] *See* Charles A. Sullivan, *Disparate Impact: Looking Past the* Desert Palace *Mirage*, 47 WM . & MARY L. REV. 911, 969-71 (2005); *accord* Barocas & Selbst, *supra* note 3, at 701 ("Where there is no discriminatory intent, disparate impact doctrine should be better suited to finding liability for discrimination [than disparate treatment]."). *But see* Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 767-68 (2006) ("[M]uch of the battle to remedy discrimination was lost when we moved away from the focus on intent.").

social action.[119] It also governs when someone has an illicit motive or seeks to intentionally discriminate in a systematic way.[120] Thus, we return to the paradox described above. It seems that an employer is trapped between using facially neutral systems that reflect systemic and historically conditioned, biased results, and facing lawsuits for using aware systems to mitigate such effects. The law, however, allows for and supports more subtle outcomes.

B. Reading Ricci *Properly*

Ricci v. DeStefano has generated legal confusion and debate about what can or cannot be done to address discrimination in hiring practices.[121] A close look at the facts and the decision reveals that rather than prohibiting action, *Ricci* provides a roadmap about methods to develop a non-discriminatory employment test and when an employer is allowed to alter a test to account for potentially discriminatory outcomes. Once understood as a roadmap, *Ricci* enables algorithmic approaches to employment practices, rather than being a roadblock.

In *Ricci*, the City of New Haven had developed a test for firefighter promotion with the help and validation of experts. When administered, 77 people took the lieutenant exam, "43 whites, 19 [B]lacks, and 15 Hispanics. Of those, 34 candidates passed, 25 whites, 6 [B]lacks, and 3

[119] *See, e.g.*, Int'l Brotherhood of Teamsters v. United States, 431 U.S. 324, 335 (1997) (employer who "regularly and purposefully treated Negroes and Spanish-surnamed Americans less favorably than white persons. . . . [by] refus[ing] to recruit, hire, transfer, or promote minority group members on an equal basis with white people, particularly with respect to line-driving positions" had engaged in disparate treatment); 42 U.S.C. § 2000e-2(a)(1) (2018) (prohibiting discrimination in employment).

[120] *See, e.g.*, McMullen v. Warner, 416 F. Supp. 1163, 1166 (D.D.C. 1976) ("racially motivated" decision was violation of Title VII); *accord* FTC, *supra* note 117, at 18 ("Systemic disparate treatment occurs when an entity engages in a pattern or practice of differential treatment on a prohibited basis."); Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1351 (2010) ("[When] the discrimination is intentional, [such] discrimination is *called* 'disparate treatment.'"); *cf.* FTC, *supra* note 117, at 18 ("[A] lender cannot refuse to lend to single persons or offer less favorable terms to them than married persons even if big data analytics show that single persons are less likely to repay loans than married persons.").

[121] *See* Barocas & Selbst, *supra* note 3, at 725-26; Bent, *supra* note 17, at 826-28; Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 925-26 (2017) [hereinafter *Data-Driven*]; Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 692-95 (2017) (observing that *Ricci* may raise "legal difficulties with correcting discriminatory algorithms ex post").

Hispanics."[122] Forty-one people took the captain's exam — "25 whites, 8 [B]lacks, and 8 Hispanics. Of those, 22 candidates passed — 16 whites, 3 [B]lacks, and 3 Hispanics."[123] Despite the experts' opinions and validations of the test, the City rejected the results because the pass rate caused the city to believe it might be sued for disparate impact.[124] The Supreme Court did not allow this after-the-fact change, because New Haven's actions relied on race (the race of those who passed the test) to reject the results. In that sense, New Haven engaged in disparate treatment. Thus, it may appear that an entity cannot account for and alter employment practices when there is evidence of potential disparate impact in the entity's practices, because such changes will necessarily be disparate treatment.[125] That is incorrect.[126]

As the Supreme Court put it, not allowing an entity to account for race to avoid disparate impact liability "even if the employer knows its practice violates the disparate-impact provision," is contrary to "Congress's intent that "voluntary compliance" be "the preferred means of achieving the objectives of Title VII."[127] This rule does not, however, mean an entity can simply assert there has been a history of past discrimination and claim there is a need to throw out a practice, because that might lead to "an unyielding racial quota."[128] As stated above, the entity has to show why the change is needed in light of the goals of Title VII. In addition, the timing of when an entity makes changes matters.

The way the test was developed and administered by New Haven in *Ricci* doomed the City's decision to reject the test's outcomes. First, New Haven began well by hiring experts to design a likely valid test. The city spent $100,000 on outside experts who designed entry and promotional tests for fire departments.[129] The hired firm conducted interviews, went on ride-alongs, interviewed incumbents at the promotional level for which applicants were being tested, and designed "job-analysis questionnaires and administered them to most of the incumbent battalion chiefs, captains, and lieutenants in the Department."[130] As the Supreme Court noted, "At

[122] Ricci v. DeStefano, 557 U.S. 557, 566 (2009).

[123] *Id.*

[124] *Id.* at 566-74 (discussing numerous meetings and steps to validate the results and the decision to reject the results nonetheless).

[125] *See* Barocas & Selbst, *supra* note 3, at 725-26.

[126] *See* Kim, *Data-Driven*, *supra* note 121, at 925-26.

[127] *Ricci*, 557 U.S. at 580-81.

[128] *Id.* at 583.

[129] *Id.* at 564.

[130] *Id.* at 564-65.

every stage of the job analyses, IOS [the company that developed the test], by deliberate choice, oversampled minority firefighters to ensure that the results—which IOS would use to develop the examinations—would not unintentionally favor white candidates."[131] Second, once the test was approved by the city, New Haven set a 3-month study period and gave candidates a study guide which included the "source material for the questions, [and] the specific chapters from which the questions were taken."[132]

The city's *ex post* actions were the problem, not the city's *ex ante* design steps. The Court rejected "invalidating the test results" after the fact without "a strong basis in evidence of an impermissible disparate impact."[133] As one scholar has pointed out, the ex-post rejection of the results created "visible victims" — that is, those who studied for the test passed and whose hard work was discarded.[134] Once the test had been given, the city needed strong evidence that the test would be invalidated if the city were sued for disparate impact and lost, because otherwise those who had passed would be harmed. The Court did not see such evidence and so did not allow the city to reject the results.

The issues around testing and the logic of *Ricci* aids understanding what one can do with algorithmic employment processes. Tests, such as the one in *Ricci*, are different than the sort of testing that occurs when assessing a pool of candidates based on a predictive algorithm. *Ricci* involved a test for which test-takers had prepared, including spending money on test preparation. As Professor Kim explains, it is a mistake to think that the tests at issue in employment cases and covered under statute address the issues raised with data and algorithms.[135] Insofar as data and algorithms test, they are not "'ability tests' because they do not actually test ability—rather, they identify behavioral markers that appear to correlate with on-the-job success."[136] As a specific example of Professor Kim's point, a resume screening is not an ability test, and that difference matters.[137]

[131] *Id.* at 565.

[132] *Id.*

[133] *Id.* at 585.

[134] *See* Primus, *supra* note 120, at 1345.

[135] *See* Kim, *Data-Driven*, *supra* note 121, at 908-09; *accord* Bent, *supra* note 17, at 842.

[136] *See* Kim, *Data-Driven*, *supra* note 121, at 908.

[137] The EEOC requires validation for "tests and other selection procedures which are used as a basis for any employment decision." *See* 29 C.F.R. § 1607.2 B (2022). Employment decisions include "hiring, promotion, demotion," and as discussed above, algorithms are used for such decisions. *Id.*; *see* 29 C.F.R. § 1607.2 C (2022) ("These guidelines apply only to selection procedures which are used as a basis for making

An entity using a resume screening system may design and test the system *ex ante* before using it far more easily than the sort of test at issue in *Ricci*. *Ex ante* action is quite different than the facts that caused problems in *Ricci*. Unlike *Ricci*, where applicants were seen as having an expectation that a potentially valid test for which they could study would be accepted, designing and using a resume screening algorithm occurs at an earlier stage of the hiring process where no hiring or promotion decision is made.[138] As Professor Kim points out, Title VII is about adverse employment *actions* such as denying a promotion as happened in *Ricci*.[139] In contrast, using an algorithm to screen resumes does not violate the law because "there has been no adverse employment action. No employee has been deprived of a job to which he is entitled because no employee has any right or legitimate expectation that an employer will use any particular model."[140] You cannot argue that a system had to select for you.

Even if one wanted to treat a screening system as the sort of test at issue in *Ricci*, *Ricci* embraces the sort of *ex ante* design steps that can and should go into building a screening system.[141] In designing a resume screening system, one might take proactive (e.g., race-aware) measures to avoid unfair or discriminatory outcomes, including making adjustments during the "training" of the algorithm.[142] These steps are analogous to the design steps — such as making overt choices and oversampling at every stage to ensure that the test did "not unintentionally favor white candidates" — taken by New Haven and of which the Supreme Court wrote with approval.[143]

Recall that one of the goals of Title VII is to reduce, if not eliminate, "unnecessary barriers to employment."[144] Thus, the *Ricci* Court did not "question an employer's affirmative efforts to ensure that all groups have a fair opportunity" at a given stage of the hiring process.[145] The key point is that an employer is allowed to examine "how to design . . . [a] practice

employment decisions. . . . but recruitment practices are not considered by these guidelines to be selection procedures."); *supra* Figure 1; *supra* notes 89–91 and accompanying text.

[138] Kim, *Data-Driven*, *supra* note 121, at 930.

[139] *Id.*

[140] *Id.*

[141] Ricci v. DeStefano, 557 U.S. 557, 565 (2009).

[142] *See* Guillaume Lemaître, Fernando Nogueira & Christos K. Aridas, *Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*, 18 J. MACH. LEARNING RSCH., 1, 3-4 (2017).

[143] *Ricci*, 557 U.S. at 565.

[144] Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

[145] *Ricci*, 557 U.S. at 585.

in order to provide a fair opportunity for all individuals, regardless of their race" before deploying it."[146]

Returning to our specific example, if an employer finds that women and minorities are rarely interviewed and further finds that resume screening to date has not selected, women and minorities for interviews, an employer should be able to design and test a bias-aware algorithm as part of voluntary compliance to avoid disparate impact. This possibility creates the need to address the validity of the new practice and what is allowed in its design.

III. BIAS MITIGATION AND ITS LIMITS: A COMPUTER SCIENCE APPROACH

When designing a decision-making algorithm, can we control for bias in internal data and thus avoid Amazon's situation discussed above? If using external data, what steps can be taken to control for historic, economic, and/or social factors that are known to skew seemingly objective metrics such as the SAT? Once one addresses biases in data, can one still have an algorithm that works well? In our example, can one control for bias and still screen resumes so that viable candidates are interviewed? And, once one finds a method to address these questions, will the chosen method fit within legal rules? Answering these questions requires understanding computer science approaches to mitigating bias.

A. Bias Mitigation: The Basics

At a general level, algorithmic bias mitigation tries to answer this question: how can one design an algorithm which performs well despite uncertainties about candidates' qualifications *and* satisfies a particular notion of fairness?[147] A variety of algorithmic techniques have been

[146] *Id.*

[147] *See, e.g.*, Corbett-Davies et al., *supra* note 73 (discussing tradeoffs in risk assessment and satisfying a particular definition of fairness); Hilke Schellman, *Auditors Are Testing Hiring Algorithms for Bias, but There's No Easy Fix*, MIT TECH. REV. (Feb. 11, 2021), https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/ [https://perma.cc/8JUA-8QRN] (noting a system may predict success better for one group such as men as compared to women). The theory of fairness in algorithmic decision-making has developed rapidly in recent years. Numerous notions of fairness have been proposed, from statistical notions seeking to mitigate disparities in statistical quantities (selection rates, FPRs, etc.) across demographic groups, to notions seeking to ensure similar treatment to similar individuals, to notions based on a causal understanding of candidate qualities. For an overview, see SOLON

proposed for coping with biased data and improving fairness. One can use pre-processing techniques which involve modifying data before feeding it to an algorithm.[148] One can use in-processing techniques, which modify the algorithm itself.[149] And one can use post-processing techniques, which modify decisions made by an algorithm after the fact.[150] No matter the stage, current computer science literature highlights that merely scrubbing protected class information from an application may not help mitigate existing biases,[151] and that algorithms have to use protected information to fix existing biases in data.[152]

One prevalent approach falls under the pre-processing category and includes iteratively removing data that is correlated with protected information, or otherwise transforming the data so that protected information cannot be recovered.[153] The idea behind this approach is that if protected information is not recoverable from data, then decisions made by a selection algorithm will be naturally fair, even if the algorithm is blind (i.e., does not use protected information). One issue is that this approach could remove highly predictive information. If the only information available was a very accurate test score, and this test score was correlated with protected information, then such an algorithm would either remove this important information or fail to make group-specific distributions indistinguishable.

BAROCAS, MORITZ HARDT & ARVIND NARAYANAN, FAIRNESS AND MACHINE LEARNING: LIMITATIONS AND OPPORTUNITIES (2018).

148 Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton & Derek Roth, *A Comparative Study of Fairness-Enhancing Interventions in Machine Learning*, 2019 CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 329, 330.

149 Houser, *supra* note 49, at 340 n.300; *see* Richard Zemel, Yu (Ledell) Wu, Kevin Swersky, Toniann Pitassi & Cynthia Dwork, *Learning Fair Representations*, 28 PROC. MACH. LEARNING RES. 1, 1 (2013), https://www.cs.toronto.edu/~toni/Papers/icml-final.pdf [https://perma.cc/W4YV-CMB3].

150 Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, 30 CONF. ON NEURAL INFO. PROCESSING SYS. 1, 3 (2016); Faisal Kamiran, Toon Calders & Mykola Pechenizkiy, *Discrimination Aware Decision Tree Learning*, 2010 IEEE INT'L CONF. ON DATA MINING 869, 871-72.

151 Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi & Adam Tauman Kalai, *Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting*, 2019 CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 120, 121.

152 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel, *Fairness Through Awareness*, 2012 PROC. 3RD INNOVATIONS IN THEORETICAL COMPUT. SCI. CONF. 214, 218, 226.

153 *See* Zemel et al., *supra* note 149, at 2-3.

Thus, imagine you are trying to address gender bias in an evaluation metric. You know that bias in evaluations can render bias-agnostic methods suboptimal.[154] Perhaps you think that using a constraint such as demographic parity — proportional selection from different demographic groups — will solve the problem. Although selection will be equal across the demographic groups, that technique can hinder performance in some cases,[155] which points to potential trade-offs between bias mitigation and quality of selections. Theoretical work has shown that in general settings, active use of demographic information is necessary in achieving certain notions of fairness.[156] In essence, to mitigate bias, we need to know who is being harmed by the bias; without demographic information (or some proxy thereof), we cannot identify who is being harmed.

There are three lessons from these approaches to bias mitigation. First, there are inconsistencies in measurement of the ability of candidates. Second, mitigating the impact of such inconsistencies is an instance-specific endeavor. Third, no cure-all exists.

The lessons point to multiple issues in bias-mitigation. First, the assumptions on bias are difficult to justify empirically, as sought qualities (e.g., ability) cannot be captured by a single number and are especially difficult to measure for candidates who have not been hired yet. Second, it is difficult to assess bias-mitigation techniques for a similar reason: if one does not know the ground truth, then it is hard to quantify how good any decision is. Despite these issues, theoretical work offers ways to mitigate bias under specific mathematical assumptions. Sections B and C explain some dominant approaches, their limits, and introduces a newer approach — the partially order set or poset approach — which opens the door to creating legally allowed, bias-aware algorithms.

---

[154] *See, e.g.*, Edelman et al., *supra* note 99, at 7 (discussing how implicit bias affects responses to Airbnb inquiries); John M. Kleinberg & Manish Raghavan, *Selection Problems in the Presence of Implicit Bias*, 2018 9TH INNOVATIONS IN THEORETICAL COMPUT. SCI. CONF. 1 (In this work, Kleinberg and Raghavan demonstrate that under certain modeling assumptions, the use of the Rooney Rule improved the quality of selections in the presence of group-specific bias in scores.); Jad Salem & Swati Gupta, *Closing the Gap: Group-Aware Parallelization for Online Selection of Candidates with Biased Evaluations*, INT'L CONF. ON WEB & INTERNET ECON. 1, 2 (2020) (under minor revision at Management Science, 2022) (discussing how implicit bias may factor into hiring decision-making).

[155] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh & Jun Sakuma, *Fairness-Aware Classifier with Prejudice Remover Regularizer*, 2012 JOINT EUR. CONF. ON MACH. LEARNING & KNOWLEDGE DISCOVERY IN DATABASES 35, 46.

[156] *See* Dwork et al., *supra* note 152, at 214.

B. Multiple Evaluation Metrics and Group Bias

Recall that a hiring algorithm will likely consider more than one evaluation metric; it will assess not just a GPA but SAT, class rank, school, etc. Using more than one metric may seem to improve evaluation; but given that each metric can have bias, using multiple metrics can increase the difficulty in assessing the accuracy of a ranking. Given these issues, attempts have been made to address miscalibration of evaluations between multiple evaluators,[157] and techniques have been developed for cases where some information is known about how biased each evaluator is in each evaluation.[158] In general, mathematical techniques can be developed if some assumptions on bias are made. Certain coarse sources of bias seem to be prevalent across demographic groups, and algorithms can be designed with these in mind. In this approach — called statistical or demographic parity — the system is designed so that the results for a protected class mirror the "proportion of the population as a whole."[159] One might say these are the first approximations to incorporate the knowledge of large trends visible broadly across demographic groups, such as are seen in SAT scores discussed earlier.[160] Addressing these coarser sources of bias from a theoretical point of view can provide insight in dealing with other forms of bias.

What if you suspect that there are important differences within a group, such as whether group members lack a certain level of education? A recent mathematical model capturing disparities in errors in testing between groups is what is called the *group model of bias*. Christine Wennerås and Agnes Wold's book, *Nepotism and Sexism in Peer-Review*,[161] provides the empirical work on which the model is based. Kleinberg and Raghavan

157 Jinyang Wang & Nihar B. Shah, *Your 2 Is My 1, Your 3 Is My 9: Handling Arbitrary Miscalibrations in Ratings*, 2019 PROC. 18TH INT'L CONF. ON AUTONOMOUS AGENTS AND MULTIAGENT SYS. 864, 865.

158 Jingyang Wang, Ivan Stelmakh, Yuting Wei & Nihar B. Shah, *Debiasing Evaluations that Are Biased by Evaluations*, 35 PROC. AAAI CONF. ON A.I. 10120, 10121 (2021).

159 Zemel et al., *supra* note 149, at 1-2 (In their approach, data is transformed in a way that "*lose*[s] *any information that can identify whether the person belongs to the protected subgroup*."); *accord* Bent, *supra* note 17, at 817 ("The simplest example of a group-fairness approach is a 'demographic parity' or 'statistical parity' approach. At its most restrictive, this would require that the predicted target variable success rates (good employee) be equal for both groups.").

160 *See supra* notes 108–12 and accompanying text.

161 Christine Wennerås & Agnes Wold, *Nepotism and Sexism in Peer-Review*, 387 NATURE 341, 342 (1997).

introduced the model in the context of offline selection.[162] Other scholars have extended the model for other selection contexts.[163] Unlike demographic parity, the group model engages with differences among group members. The model assumes that bias is somewhat consistent within each demographic group, and thus evaluations offer more accurate rankings within each group but may offer inaccurate rankings between members of different groups. As such, this approach begins to incorporate potential structural differences for groups such as history of racism or denied access to credit as they affect a particular group.

Under the group model of bias, inaccuracies in rankings between candidates in different groups are deemed unreliable comparisons, and so one can only make reliable comparisons between candidates in the same group. For example, imagine a test for scholars scored on a percentage basis, and the group model of bias with respect to gender is assumed. Bias can be present and affect any group, so the test will likely have biases for each group. The model assumes that bias within a group *uniformly* affects members of that group (e.g., evaluations of female scholars may be underestimated by a factor of 0.9); since the effect is assumed to be uniform, comparisons between candidates in the same group are unaffected by the bias. The model will allow for a more robust understanding of the differences between white male scholar Adam and Latino scholar John, where Adam scores 90% and John scores 83% and find that Adam did better than John. If, however, one looks at Jennifer, a Latina scholar who scored an 85% and Adam, with 90%, one cannot compare the two, because the model assumes a bias based on gender.

This model is at the same time appealing and dissatisfying in its simplicity. It is appealing in the sense that the model sheds light on best practices *when the data is biased consistently for certain groups*. That consistency indicates that information about group membership alone allows selection algorithms to reduce bias in selections. It is dissatisfying, however, in its coarseness. It will assume that all within a group are the same. It ignores intra-group differences in testing/evaluation errors and ignores any potential comparisons between groups. In the example above, there may be racial biases within each gender group that are not accounted.

[162] *See* Kleinberg & Raghavan, *supra* note 154, at 4.

[163] *See, e.g.*, Avrim Blum & Kevin Stangl, *Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?*, 156 LIPICS 3:1 (2020) (extending the model to labeling bias from underrepresentation bias); Faenza et al., *supra* note 112 (extending the model to school admissions); Salem & Gupta, *supra* note 154 (extending the model by allowing members of the same group to be incomparable and allowing members of different groups to be comparable).

Under the group model of bias with respect to gender, Adam and John can be ranked, whereas under the group model of bias with respect to race, they cannot. Further, considering more granular groups (e.g., one group could be males, with low-income, aged 35-39) is not practical solution, as group sizes may become very small, thus inhibiting the employer's ability to make informed decisions. The group model of bias is therefore limited in its ability to provide individualized treatment of candidates.

To address multiple, possibly dependent variables, follow-up work proposed a multiplicative model of bias in the context of rankings, wherein candidates in the intersection of different groups face a consistently higher bias.[164] Tying this to our example above, where gender and race are considered, we might assume that women face more bias than men, and Latinx people face more bias than white people. Thus, the Latina scholar Jennifer (with score 85%) would be ranked above the Latino scholar John (with score 83%); we can make this comparison under the multiplicative model despite the scholars' different genders since women were assumed to face greater bias. This model thus allows for more comparisons to be made than the group model with the same demographic breakdown would allow. This approach, however, again equalizes the amount of bias within each smallest intersectional group (e.g., male, white, and age above 45 or lesbian, Asian, aged 39). It may not be okay to equalize the experience of every male, white person above the age of 45 or of every lesbian, Asian, under age 50. The underlying problem though with this work is the assumption of group membership, which may not even be accurate in practice. Indeed, whether a Chinese Asian, an Indian Asian, and a Filipino Asian should be treated the same seems unlikely.[165]

---

[164] L. Elisa Celis, Anay Mehrotra & Nisheeth K. Vishnoi, *Interventions for Ranking in the Presence of Implicit Bias*, 2020 CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 369, 373.

[165] *See* Anna Purna Kambhampaty, *At Census Time, Asian Americans Again Confront the Question of Who 'Counts' as Asian. Here's How the Answer Got So Complicated*, TIME (Mar. 12, 2020, 12:15 PM EDT), https://time.com/5800209/asian-american-census/ [https://perma.cc/4R9S-DWWR] (tracing the history of the term "Asian" and differences in meaning depending on the country where it is used); Li Zhou, *The Inadequacy of the Term "Asian American,"* VOX (May 5, 2021, 10:10 AM EDT), https://www.vox.com/identities/22380197/asian-american-pacific-islander-aapi-heritage-anti-asian-hate-attacks [https://perma.cc/W5J5-ADBB].

C. New Frontiers in Algorithms for DEI: The Partial Ranking, or Poset, Approach to Bias Mitigation

Coping with uncertainties in data is a fundamental problem in applicant screening systems, as well as in data-driven decision-making more generally. Because the group model and related refinements of it fail to capture the complexities of errors in broad settings, other approaches are needed. The *poset approach* for applicant-screening in the face of uncertainty has emerged recently in the computer science literature and offers a new way to tackle the problems of uncertainty and fairness.[166] By extension, the poset approach provides new ways to account for bias.

1. Fundamentals of Posets and the Poset Approach

Consider the following scenario: there are three candidates A, B, and C competing for two interview slots. A has an ability score of 82; B, 68; and C, 67. You know that the ability score is a strong predictor of job performance. You also know the score is accurate up to 3 points. As such, there is a significant chance that C is a better candidate than B. The problem is that if you simply select based on highest scores, you always select A and B, despite knowing that the scores of B and C are within the accuracy range. The core idea behind the poset approach is that the latter approach is unfair to C, or more generally, that ignoring uncertainty can result in unfair decisions. In other words:

> Some applicants, due to individual experiences or lack of historic data, cannot be reliably ranked. The solution need not involve producing a (possibly inaccurate) ranking. Instead, allowing for partial rankings can itself open the door to fairer decisions.

A particular aspect of this method shows that pairs of candidates can be incomparable, given enough uncertainty in the data.[167]

The poset approach, which we explain in more detail below, makes use of a mathematical structure called a *partially ordered set*, or *poset*, which can be used to encode uncertainty in ordinal information. Consider, for example, a set $S_1=\{4,2,5\}$ of true hirability of three candidates (which is often not observable in practice). This set is called *totally ordered* since any pair of the scores can be ordered (i.e., ranked) with respect to the relation "$\leq$". In other words, we can rank the scores: $2 \leq 4 \leq 5$, thus inducing an order

[166] Salem & Gupta, *supra* note 154, at 11-12.

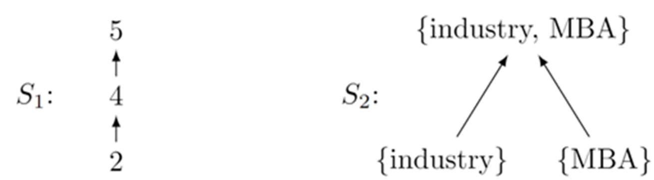
[167] *Cf.* Barocas & Selbst, *supra* note 3, at 685-87 (noting datasets for a given class may be less robust than for another class).

among the candidates. The problem is that in practice, one cannot observe directly how good a candidate might be at their job. This is where partial orders can help us.

One can think of a partial ordering as a set of comparisons, which may not cover all pairs of candidates (i.e., a total order with some comparisons missing). For example, consider a candidate A who has experience in an industry; candidate B who has experience in the industry *and* who has an MBA; and candidate C who has an MBA but no experience in the industry. Considering these traits as binary (yes/no) attributes, one can represent their qualifications as the set S_2 ={{industry}, {MBA, industry}, {MBA}}. From the given information, one might rank B above both A and C, because B is qualified with respect to both measures, and the other candidates are only qualified with respect to one. A and C might be considered *incomparable* because their qualifications are complementary. That is, the traits "in industry experience" and "MBA" both matter for the potential employer, but because A lacks an MBA and C lacks experience, we cannot assess them directly to each other.

In this case, S_2 is a poset, but not a totally ordered set. A poset is often visually depicted using its *Hasse diagram*, which is a directed graph in which edges represent orderings. For example, the Hasse diagrams for S_1 and S_2 are as follows:

Figure 1. Hasse diagrams for S_1 and S_2.



Note that Hasse diagrams omit redundant edges: even though $2 \leq 5$, the edge $2 \rightarrow 5$ is not included, since it is implied by the edges $2 \rightarrow 4$ and $4 \rightarrow 5$.

The *poset approach* is the process of (1) forming a partial ranking (i.e., a partial order) of the candidate pool based on uncertainties, inaccuracies, or biases in data, and (2) making selections based on this poset. By making selection decisions in this way, one can concretely take uncertainty into account and, for example, avoid routinely harming candidate C in the S_2 example above.

The approach opens the door to addressing some core concerns regarding using data and software. One concern is that the data used to train a system can "[s]ystematically disadvantage those who are under- or over-represented in the dataset."[168] A related concern is that if a data analyst chooses only certain features or attributes to fuel the system, those choices again can leave out certain groups.[169] For example, an employer may choose where someone went to school as a feature and that may be useful because the data cost less than developing a truly rich, granular account of each person under consideration.[170] Thus that approach will leave out those who may deserve consideration.

There are unstated assumptions in the critique. First, as discussed above, the concerns track problems with poor data science practices and the assumption that people will use such poor practices rather than known best practices. That is a human issue; not a software one. Recall that both the Amazon resume and Boston pothole examples showed responsible management of data science rather than allowing undesired outcomes to take hold. Second, the views assume that one can somehow perfectly manage either over or under representation in data. Third, it assumes that one can access ideal feature data but choose not to use it. All these assumptions miss a larger, crucial point.

Even with best practices regarding data and feature selection, uncertainty is a reality that must be addressed. The poset approach's power is that it explicitly acknowledges issues in data and feature selection,

---

[168] *Id.* at 681.

[169] *Id.* at 689.

[170] *Id.* ("Obtaining information that is sufficiently rich to permit precise distinctions can be expensive. Even marginal improvements in accuracy may come at significant practical costs and may justify a less granular and encompassing analysis."). Ironically, this point suggests the need for reduced privacy because less privacy implies more data, which is what Barocas and Selbst seem to say is needed to solve the problems of poor datasets. *See, e.g.*, FRANK WEBSTER, THEORIES OF THE INFORMATION SOCIETY 208 (John Urry ed., 3d ed. 2006) ("If we are going to respect and support the individuality of members, then a requisite may be that we know a great deal about them."); Desai, *supra* note 27, at 557-59 (describing how individuation, having more knowledge about someone, allows for recognizing individuality). A further irony is that the core idea of the article — "big data discriminates" — flows from the premise of large so-called "big data" datasets, and so for the critique to hold, it also assumes that data analysts are not using big data and not adhering to the claims of big data advocates such as Mayer-Schonberger and Cukier who argue that the cost for such granularity is quite low. VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 47-48 (2013) (claiming that the era of Big Data is a revolution that uses large, messy datasets and that knowing causes no longer matters, because the data set is so large that errors having to do with sampling and other problems in the data are overcome); *accord* RACHEL SCHUTT & CATHY O'NEIL, DOING DATA SCIENCE: STRAIGHT TALK FROM THE FRONTLINE 24-26 (2013).

accounts for such issues, and enables bias mitigation in cases where the evaluation metric is biased against a certain group. If a group is underrepresented in training data and experiences large errors in the resulting ML model, the poset approach can confer benefit of the doubt to those underrepresented candidates.

2. Posets: Examples in Action

One example of poset use illustrates how they can aid in accounting for bias. The example is for sexual orientation bias but can apply to age, gender, race, etc. The Subsection then explains subtle aspects of the approach. The other example shows how posets help assess candidates when using more than one feature.

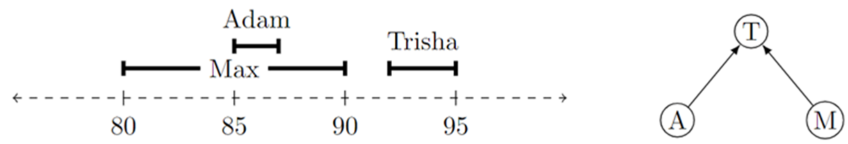
a. Using Posets to Account for Sexual Orientation Bias

The poset approach offers a way to account for uncertainties while also avoiding prohibiting discrimination such as based on sexual orientation.[171] Using the poset approach, one may incorporate demographic context of the candidates and quantify uncertainty in their evaluations.

Suppose that in a training dataset, gay male candidates are underrepresented and consequently have high variance in errors in the prediction model. Suppose there is a candidate Max with a predicted score of 85, Adam with a predicted score of 86 and say Trisha with a predicted score of 92. Accounting for uncertainties due to lesser data on candidates like Max (i.e., Max is a gay male), a poset approach would assign Max a wide score range of 80-90% (i.e., Max's true score lies in [80,90]). A poset approach might find that there is lesser error or uncertainty about the score of candidates like Adam, and assign Adam a score range of 85-87%, and similarly, Trisha might get assigned a score range of 92-95% (see Fig. 3).

[171] As recently as June 15, 2020, the Supreme Court of the United States ruled that the Title VII of the Civil Rights Act prohibits discrimination on the basis of sexual orientation and gender identity. Bostock v. Clayton Cnty., 140 S. Ct. 1731, 1737 (2020) ("Today, we must decide whether an employer can fire someone simply for being homosexual or transgender. The answer is clear. An employer who fires an individual for being homosexual or transgender fires that person for traits or actions it would not have questioned in members of a different sex. Sex plays a necessary and undisguisable role in the decision, exactly what Title VII forbids.").

Figure 2. Score ranges and induced partial ranking for three candidates.



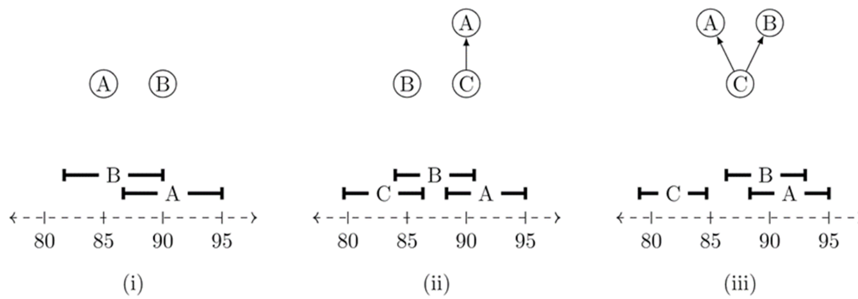
Using only the score ranges to compare candidates, a manager can decide that they will only make recommendations for hiring if a candidate's score range does not overlap with the others and is better (i.e., lower boundary of the score range is higher than the higher boundaries of the other candidates). Given this data on score ranges, the manager may decide to hire Trisha, but may not be able to compare Adam and Max as their score ranges overlap. In such a case, a manager might either invest more resources (such as extra interviews or internships) to get more information about Max and Adam and decrease the width of their score ranges OR select one of them at random. In either case, using the uncertainty around the scores, the manager can avert a situation where Max has a score of 85, Adam has a score of 86 and the manager blindly selects Adam due to a difference of a contentious 1 point. Consistently selecting based on a slight difference in scores that are within the range of uncertainty can create a built-in headwind in the hiring pipeline. The poset approach therefore allows for more individualized treatment of inconsistencies in data processed as compared to the group model, since candidates in differing groups can still be compared if their confidence intervals do not overlap.

Note that in the example above, score ranges were used to account for differing amounts of uncertainty based on sexual orientation, but we glossed over *how* the ranges were constructed. Two important questions for an employer wishing to protect a group (say, gay males) using the poset approach are: "do I have access to the relevant protected information?" and "do I want to use the relevant protected information in the construction of score ranges?" If the answer to both questions is yes, then the employer may be able to directly observe error rates for each group and choose score ranges accordingly. However, the relevant protected information (say, the sexual orientation of the applicant) may not be available, or the employer may want to avoid using the information to avoid scrutiny. In such cases, there are still ways to account for differing levels of uncertainty for different groups. Unsupervised methods can be used to cluster similar individuals together, and cluster-specific errors can be used to construct score ranges. Importantly, however, the clusters formed may not recover sexual orientation, or any other group which the employer wishes to protect. This illustrates the trade-off between fairness and blindness discussed in Section III.A.

b. Subtleties in the Poset Approach

Above, we discussed how one can form a partial ranking from confidence intervals and make selection decisions. Here, we provide several more examples to demonstrate how ordinal information can translate to treatment under the poset approach.

Figure 3: Illustration of possible confidence intervals of three candidate pools, along with their induced partial rankings.



In Figure 4(i), there are two candidates in question, A and B, and their confidence intervals overlap. In this case, the two candidates are indistinguishable in the partial ranking, so an employer using the poset approach may treat them equally.

A pair of incomparable candidates should not, however, necessarily be treated equally. For example, consider the scenario in Figure 4(ii). If incomparable candidates were to be treated equally, then candidates A and B would be treated equally, as would candidates B and C; this, however, implies that candidates A and C would be treated equally despite A being confidently better than C. Importantly, using the poset approach, one need not enforce a blanket rule such as requiring that candidates with overlapping score intervals should be treated equally,[172] and therefore, A can be given preference over C.

In Figure 4(iii), candidates A and B are incomparable, and candidate C is confidently worse than both A and B. In this case, candidates A and B

[172] *E.g.*, Matthew Joseph, Michael Kearns, Jamie Morgenstern & Aaron Roth, *Fairness in Learning: Classic and Contextual Bandits*, *in* 29 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 1, 4 (2016) (introducing a fairness constraint in which those with overlapping intervals must be treated equally).

could receive equal or similar treatment due to their similar placement in the partial ranking, and this treatment could be better than that of candidate C.

c. Using Posets to Compare Candidates Based on Two or More Features

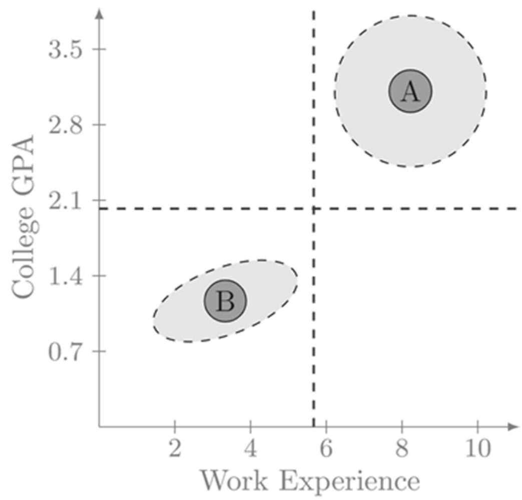
Comparing Candidates Using Two Features. Suppose that candidates are evaluated on two attributes (e.g., work experience and college GPA), and selections are to be made based on these evaluations. Note that these scores will induce two different rankings (partial or total): one with respect to work experience, and one with respect to college GPA. The complication with this scenario is that these two rankings may be contradictory: candidate A may have more work experience than B, and B may have a higher college GPA than A. In this case, the hiring manager needs a way to reconcile these two rankings.

If the hiring manager is making decisions based on point scores, without accounting for uncertainties or biases, then one solution is to weight the two attributes (say, 60% work experience, 40% college GPA), thereby recovering the one-feature scenario discussed earlier. Weighting in this manner is common practice.[173] However, if one wishes to account for uncertainties, then another approach is needed.

Suppose each candidate has a raw score (x,y), where x measures their work experience and y is their college GPA. Suppose further that each candidate has a known confidence region (i.e., it is known that a candidate's true ability scores lie within the confidence region with, say, 95% confidence). See, for instance, Figure 5, which depicts raw scores and confidence regions for two candidates. The shape of this confidence region may depend on properties of the two evaluation metrics. For example, if errors by the two metrics are suspected to be independent, then the confidence regions may be rectangles formed by confidence intervals along each attribute. If there are dependencies between the two evaluations, then confidence regions may have different shapes, as in Figures 5-6.

[173] This outcome could be subconscious. Someone may rate candidates by work: a > b > c, or by GPA: c > b > a. If they have c > a > b, that person can find a weighting of the two that will give c > a > b.

Figure 4. A depiction of two candidates' scores (marked by dark gray circles) as well as confidence regions (shown in light gray). These regions are analogous to confidence intervals: with some high degree of confidence, each candidate's "true" ability lies within their confidence region.



Given confidence regions for all candidates, how can one derive partial rankings? One way is to conjoin the partial rankings along both attributes: if A has more work experience than B *with high confidence* (i.e., their confidence intervals do not overlap) and A higher college GPA than B *with high confidence*, then A is ranked above B. More formally, if for any (x_1,y_1) in the confidence region of A and any (x_2,y_2) in the confidence region of B, $x_1 > x_2$ and $y_1 > y_2$, then A is ranked above B. Figure 4 illustrates this concept: then vertical red line shows that A has better work experience than B with high confidence, and the horizontal red line shows that A is better than B on college GPA with high confidence. Those observations together allow us to rank A above B in the derived partial ranking. A third way to explain this process is as follows: from each two-dimensional confidence region, we can infer two one-dimensional confidence intervals by ignoring the other attribute (see Fig. 6, top right, which has these confidence intervals projected on the two axes). From these two sets of confidence intervals, we can construct two partial orders, one for each attribute (see Fig. 6, middle right). Finally, we say that A is ranked above B if this is the case in *both* partial orders. We further illustrate this concept using a more grounded example next.

Example. Suppose that three candidates are to be selected based on two attributes: work experience and college GPA. You have set cutoffs for each of these attributes and only wish to select candidates exceeding each cutoff. See Figure 6 (top) for a depiction of the candidate pool, where each color represents a particular demographic group. Let the colored areas around each candidate node represent a "confidence region;" i.e., with some high degree of confidence, the candidate's latent ability lies in the drawn region. We can infer partial rankings from these confidence regions as discussed above: if the confidence region of candidate A is strictly above and to the right of the confidence region of candidate B, then A is ranked above B.

Using only raw scores, only the two blue candidates meet the cutoffs (Fig. 6, top left). However, taking confidence regions into account, we see that the two green candidates might meet the cutoffs as well (Fig. 6, top right). How, then, should one choose three candidates among the green and blue ones? One way to do so is to look at the partial ranking induced by the confidence regions (shown using arrows in Fig. 6, bottom right). In this partial ranking, there are three candidates who are maximally ranked (i.e., are not ranked below any other candidates): the two blue candidates and the right-most green candidate. This is one justification for selecting candidates 1, 2, and 4. However, additional information and hiring preferences can come into play here as well — partial rankings are meant to guide selection decisions, not to be the be-all-end-all as that would resemble blind hiring.

It is important to note that the poset approach is not the only way to make selection decisions from multi-dimensional data. Given the complexity and novelty of the poset approach, it is worth considering these alternatives. For example, one could extract a single score for each candidate using a *weighting* of the features. For example, one could weight GPA by 0.5 and work experience by 0.5, or weight GPA by 0.75 and work experience by 0.25 and add the scores together to extract a single score for each candidate (Fig. 6, bottom left). Having given each candidate a single score, the candidates can be fully ranked. Indeed, hiring committee members may implicitly evaluate candidates in this manner due to their assessment of the relative importance of the features.[174] Although this

[174] This observation raises a question regarding consistency of committee members' rankings: if each committee member reduces candidates to a single score using an implicit weighting of features, is it possible to identify inconsistencies from the given scores or rankings of candidates? For example, suppose candidate A has a score of 2 for GPA and 1 for work experience, candidate B has a score of 1 for GPA and 2 for work experience, and candidate C has a score of 3 for GPA and 0 for work experience. If a committee member ranks these candidates as A > B > C, then there is certainly an inconsistency. Since A > B, the member's weighting for GPA is larger than their weighting for work experience, but

weighting method is unambiguously simpler than the poset approach, any chosen weighting must be validated as a ranking method, as discussed in Section IV.B.1, and it is not necessarily the case that a valid weighting even exists. Moreover, using weightings to reduce candidates to single numbers (or even to single intervals[175]) destroys potentially useful information.[176] Adopting a poset approach can help avoid both issues.

That said, if it is known or suspected that some (unknown) weighting of features results in a valid ranking of candidates, then the poset approach can be used to make selection decisions. For example, suppose the features are GPA (0.0 to 4.0) and SAT scores (with SAT scores in percentiles), and candidates A, B, C, D, and E have (GPA, SAT) scores of (1.0, 50), (2.0, 75), (3.0, 25), (3, 75), and (4.0, 25), respectively. To weight the features on the same scale, it is useful to scale the scores to be in the same range (otherwise, a weighting of 0.5 for GPA and 0.5 for SAT would confer much more importance to SAT scores than GPA). So, scaling GPA by ¼ and SAT percentiles by 1/100, the new scores are (0.25, 0.5), (0.5, 0.75), (0.75, 0.25), (0.75,0.75), and (1,0.25). Now suppose that a valid weighting of the GPA lies in [0.6,0.8] (and so, a valid weighting of SAT scores lies in [0.2,0.4][177]). Each weighting in the given range produces some ranking of the candidates. The possible rankings in this example are A < B < C < D < E and A < C < B < E < D. Since the relative ranking of B with C differs across the two rankings, those two candidates are incomparable (and similarly for D and E). We are therefore left with the partial ranking shown in Figure 7. Using the poset approach in this way, the hiring committee might decide to treat D and E equally and to treat B and C equally.

this contradicts the ranking B > C. It is an open question how and when such inconsistencies can be inferred.

175 For example, suppose a candidate has a confidence interval of [1,3] for GPA and [6,10] for work experience. Given a weighting of 0.4 for GPA and 0.6 for work experience, one can extract a combined interval of [(0.4)(1) + (0.6)(6), (0.4)(3) + (0.6)(10)] = [4,7.2].

176 Suppose, for example, that candidate A has confidence intervals of [1,3] for GPA and [1,2] for work experience, and candidate B has confidence intervals of [2,4] for GPA and [8,10] for work experience. Since the candidates' GPA intervals overlap, they would be deemed incomparable under the poset approach. However, if we weight GPA by 0.5 and work experience by 0.5, we can extract combined intervals of [1,2.5] for candidate A and [5,7] for candidate B, thus making the candidates comparable under the single-dimensional poset approach. So, in reducing the data from two dimensions to one dimension, we lost the information that made the candidates incomparable.

177 Weightings for the purpose of ranking only matter in the relative sense. For example, a weighting of 1 for GPA and 1 for SAT produces the same ranking as a weighting of 0.5 for GPA and 0.5 for SAT. So, we assume without loss of generality that the sum of the weights must be 1.

Figure 5. A depiction of candidate scores and induced rankings with respect to two attributes: work experience and college GPA. On the left, rankings are derived from point scores, and on the right, rankings are derived from confidence regions. Scores and confidence regions are plotted in the top row; rankings with respect to each attribute in isolation are shown in the middle row; and rankings derived from both attributes (via weighting on the left, and in the manner discussed in Figure 4 on the right) are shown in the bottom row.

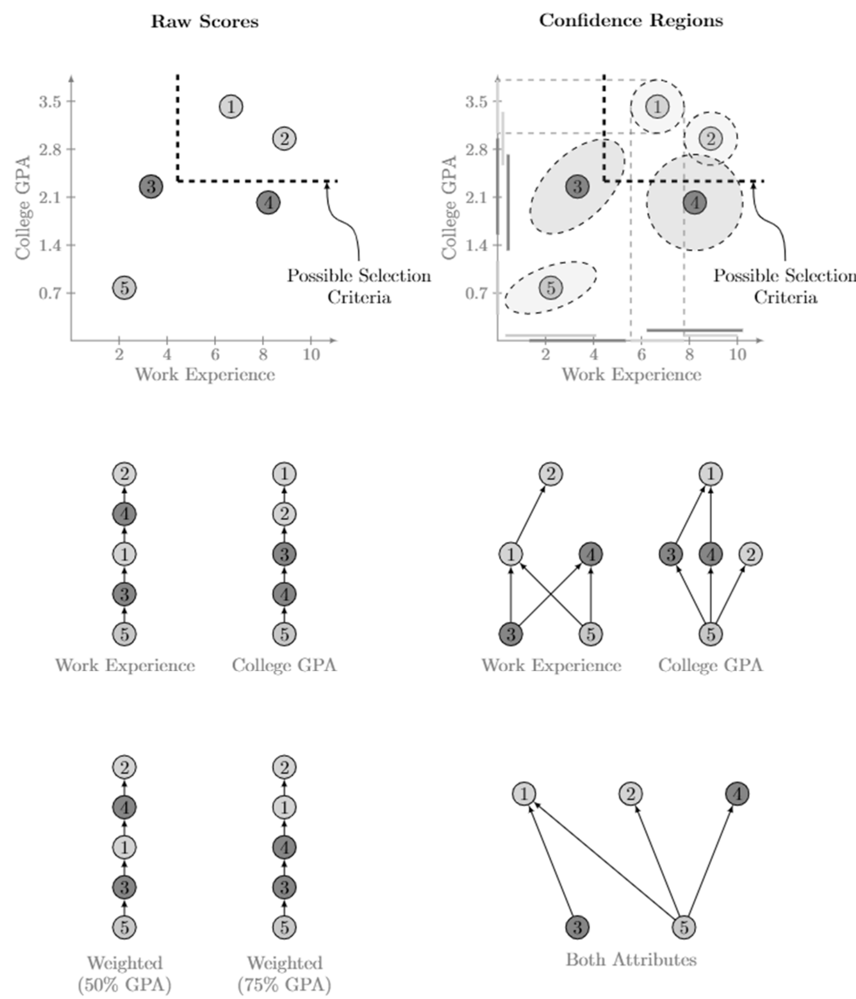
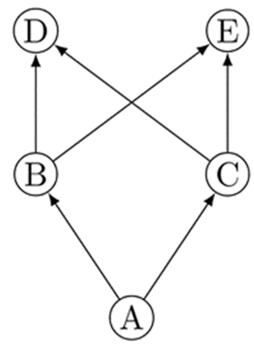


Figure 6: Partial ranking of the candidates described in III.C.2c.



3. How Posets Allow for More Individualized Assessment

The process outlined in these examples (forming score ranges/regions for each candidate and inferring comparisons therefrom) can be applied quite generally and allows for explicit treatment of bias in data. Data-driven techniques, such as estimating latent group-bias in a machine learning model, can be applied to generate these score ranges, which in turn induce a partial ranking. Such methods can be used to avoid penalizing applicants who come from underrepresented groups, who are more likely to face inaccurate evaluation via machine learning models. A recent paper by Emelianov, Gast, Gummadi, and Loiseau, shows that groups with high error variances can receive worse treatment, even if the evaluations are unbiased for all candidates, pointing to the need for interventions like the poset approach that take uncertainty into account.[178]

Simply put, the approach allows one to account for problems flowing from underrepresentation and other biases in a dataset based on acknowledging and accounting for variances and noise in data as they relate to ever finer groupings. The approach avoids the problems of pure ranking and enables more individualized assessment as compared to rank-ordering. The approach thus addresses the critique that using ML to sort candidates often leads to suboptimal or unfair treatment because of problems with poor datasets *and* offers a way to assess candidates on a more wholistic basis. Overall, ways to account for uncertainty and undesirable trends documented in machine learning models might provide a way forward for due process, while potentially improving demographic

178 Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi & Patrick Loiseau, *On Fair Selection in the Presence of Implicit Varianc*e, 2020 ACM CONF. ON ECON. & COMPUTATION 649, 658.

representation. The question now becomes which techniques pass legal muster?

IV. Legal Rules and Algorithmic Actions

How do companies hire candidates, while reconciling with the discrimination laws and biases in the hiring pipelines? This Part sets out attempts to use algorithms to navigate discrimination laws. It turns out that trade-offs in algorithmic design track tensions in legal doctrine. As discussed above, one can control for some aspects of bias, but depending on one's goal, the solution may call for a bias-aware algorithm. This Part sets out the legal and algorithmic tensions and then offers a path forward.

A. Attempts to Make Algorithmic Hiring Fit Within the Law

1. The 4/5ths Mistake

A recent approach to protecting hiring algorithms embraces a legal guide but misunderstands, and so over-estimates, the power of that guide. The guide in question is the "four-fifths rule" rule. The "four-fifths rule" comes from the EEOC's Uniform Guidelines on Employee Selection Procedures (1978) (the "EEOC Guidelines"), which provide:

> A selection rate for any race . . . which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.[179]

It is the selection rate that matters. For example, imagine an employer uses a promotion system and there were 100 female and 100 male candidates. After applying the system, only 20 females were selected while 40 males were selected. The selection rate for females would be 50% of the male rate (20/40). The 50% ratio is less than 80% or 4/5ths and so "would violate the four-fifths rule and demonstrate adverse impact."[180]

[179] 29 C.F.R. § 1607.4(D) (2022); *accord* Bradley v. City of Lynn, 443 F. Supp. 2d 145, 160 (D. Mass. 2006).

[180] *Bradley*, 443 F. Supp. 2d at 160; *see also* Ricci v. DeStefano, 557 U.S. 557, 586-87 (2009) (City of New Haven rejecting test because outcomes violated 4/5ths rule).

One might think that simply following the rule would shield an employer but that is not so. The rule is not law; it is an EEOC guideline.[181] The Supreme Court has said it is nothing more than "a rule of thumb for the courts."[182] Indeed, the Court has explicitly stated that plaintiffs challenging such a practice "must offer statistical evidence of a kind and degree sufficient to show that the practice in question has caused the exclusion of applicants for jobs or promotions because of their membership in a protected group."[183] But the "formulations" of what that proof must be "have never been framed in terms of any rigid mathematical formula."[184] Not violating the rule avoids a presumption of adverse impact, but it does not insulate an entity from challenges of discrimination.[185]

Nonetheless, perhaps because the rule appears to be the sort of specification computer scientists like — a precise numeric rule with boundaries — the rule has been used to audit and justify screening algorithms. In a recent study, the only specific public claim made by vendors of pre-employment assessments was adherence to the 4/5ths rule — outlined in the 1978 Uniform Guidelines on Employee Selection Procedures — that no pre-screening would select less than 4/5ths of candidates in a group compared to any other group.[186] One company in the AI-based hiring industry, Pymetrics, explicitly relies on the 4/5ths rule to validate its results. It even shared its data with researchers to audit the program and see whether it behaved as promised.[187] This approach fits well within technical accountability as Professors Desai and Kroll have developed the idea.[188] In simple terms, technical accountability works when an entity identifies a particular specification in an algorithm and claims to follow it.[189] Then a third party tests to see whether the specification was followed.[190] As one commenter noted, the Pymetrics

[181] 29 C.F.R. § 1607.4(D) (2022); *accord Bradley*, 443 F. Supp. 2d at 160.

[182] Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 995 n.3 (1988).

[183] *Id.* at 994.

[184] *Id.*

[185] *Id.* at 1003-04; *accord* Schellman, *supra* note 147.

[186] Manish Raghavan, Solon Barocas, Jon Kleinberg & Karen Levy, *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, 2020 CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 469, 472-73.

[187] Schellman, *supra* note 147.

[188] *See* Desai & Kroll, *supra* note 32, at 11.

[189] *Id.*

[190] *Id.*

audit may show that the company did what it said it would do but does not answer, "Are they doing the correct or right thing?"[191]

In addition, the approach is coarse as it is agnostic to quality or ability of candidates. Applying a 4/5ths rule in selection up front (e.g., as the current practice in the industry suggests)[192] does not change the perceived potential of candidates, nor account for uncertainties and biases in the data systematically. As Pauline Kim has noted, insofar as following the rule does not identify, for example, the "best qualified women," it may lead to hiring women who are not "as successful on the job."[193] Therefore following the 4/5ths rule can set up the underrepresented group's candidates for failure, and lead to resentment and enlivening of negative stereotypes.[194]

2. Other Approaches to Protect Algorithmic Hiring

Other prevalent approaches include transforming data so that protected information cannot be observed or guessed (e.g., by iteratively removing data that is correlated with protected information). This is typically done until the feature distributions for each protected class are indistinguishable.[195] The idea behind this approach is that if protected information is not recoverable from data, then decisions made by a selection algorithm will be naturally fair, even if the algorithm is "blind." One issue is that this approach could remove highly predictive information. If the only information available was a very accurate test score, and this test score was correlated with protected information, then such an algorithm would either remove this important information or fail to make group-specific distributions indistinguishable.[196]

---

[191] Schellman, *supra* note 147 (quoting computer scientist Manish Raghavan, who researches algorithmic hiring systems).

[192] Raghavan et al., *supra* note 186, at 472.

[193] Schellman, *supra* note 147.

[194] *Id.* (noting a system that does not identify "best qualified women" may lead to hiring women who are not "as successful on the job"); M.J. Fischer & D.S. Massey, *The Effects of Affirmative Action in Higher Education*, 36 SOC. SCI. RSCH. 531, 540 (2007); Madeline E. Heilman, Caryn J. Block & Peter Stathatos, *The Affirmative Action Stigma of Incompetence: Effects of Performance Information Ambiguity*, 40 ACAD. MGMT. J. 603, 611 (1997) (noting that "when [success information] was ambiguous, success information did little to attenuate negative reactions to women associated with affirmative action").

[195] Zemel et al., *supra* note 149, at 2-3.

[196] *Cf.* Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 992-93, 998-99 (1988) (noting employers should not use methods that lead to quotas and have large discretion in using methods to predict performance).

3. The Rooney Rule and the Quota Problem

Setting aside seats for interviews can become a quota and undermine efforts at diversity. Roughly twenty years ago, the National Football League ("NFL") followed the lead of Dan Rooney, owner of the Pittsburgh Steelers, and required teams interviewing for head coaching positions to interview minority candidates.[197] One problem with this approach is like the problem when one simply follows the 4/5ths Rule; people may be selected to satisfy the feeling that diversity is being attained or at least honored, but whether candidates are taken seriously is dubious. The recent lawsuit against the NFL indicates that rather than assess candidates on merits, teams have been saving a seat for a minority interview as show. The team may have already anointed a candidate but went through the motions of interviewing other candidates including minority ones. As Eric Bieniemy, offensive coordinator for the Kansas City Chiefs during the team's recent dominant and successful past few seasons, has said, "some of those guys were legitimately looking at me as a possible head coach. Other guys were just, you know, carrying out their Rooney Rule . . . obligations."[198] The NFL's chief diversity officer argues that nonetheless, the Rooney Rule has increased the number of minority candidates "in the room to compete for roles, [so] the opportunity of a diverse candidate getting hired goes up."[199] The two perspectives connect with computer science on the issues of addressing inequity.

The Rooney Rule smacks of being a quota with no substance other than saying a team interviewed a minority candidate. Quotas can, nonetheless, sometimes address bias. A recent paper by Emelianov et al. shows how quota-based systems can mitigate the effects of disparate error rates across groups.[200] Quotas, however, are non-individualistic and can be illegal.[201] In addition, processes that look like or turn into quotas end up creating a bias in interviewers who may overtly or subconsciously think that the

---

[197] Scott Neuman, *Why a 20-Year Effort by the NFL Hasn't Led to More Minorities in Top Coaching Jobs*, NPR (Feb. 3, 2022, 1:23 PM ET), https://www.npr.org/2022/02/03/1075520411/rooney-rule-nfl [https://perma.cc/9Q4S-GUA3].

[198] *Id.*

[199] *Id.* (quoting Jonathan Beane, the NFL's chief diversity and inclusion officer).

[200] *See* Emelianov, *supra* note 178 and accompanying text.

[201] *Watson*, 487 U.S. at 993, 994 n.2 (Blackmun, J., concurring in judgment) ("Preferential treatment and the use of quotas by public employers subject to Title VII can violate the Constitution . . . ." (citing Wygant v. Jackson Bd. of Educ., 476 U.S. 267 (1986), Albemarle Paper Co. v. Moody, 422 U.S. 405, 449 (1975))).

candidate is being interviewed but lacks the qualifications for the position.[202]

In contrast, using score ranges, such as the poset approach, instead of raw scores does not set up a quota system. Unlike a quota, the poset approach results in selection rates dependent on the data (e.g., rankings may depend on how represented each candidate is in the data). The key to using a bias-aware algorithm such as the poset approach of Salem and Gupta is to establish the facts and evidence of a need to address bias (i.e., inconsistencies in the data) as set forth above, and then to build a plan that assesses individuals rather than setting up a purely number-driven process with percentages (i.e., quotas)[203] for each category.[204] It creates opportunities for candidates who have more uncertainty in their evaluations, while following the popular principal of "optimism in the face of uncertainty."[205]

Even though accounting for uncertainties will often benefit the underrepresented group, this may not always be the case (see, e.g., Fig. 6, which shows that increasing lengths of confidence intervals under the poset approach can benefit or harm any group). One need not, however, discard a poset-based approach if the selection rate of an underrepresented group decreases. If the approach were properly validated, and no other less discriminatory method for accounting for the uncertainties were known, then illegal disparate impact may be less likely to be the issue. Because the poset approach may benefit or harm *any group* depending on confidence intervals, it avoids being a quota and is instead an approach that can be interrogated regarding possible discrimination problems to see whether the way it is built meets legal standards.

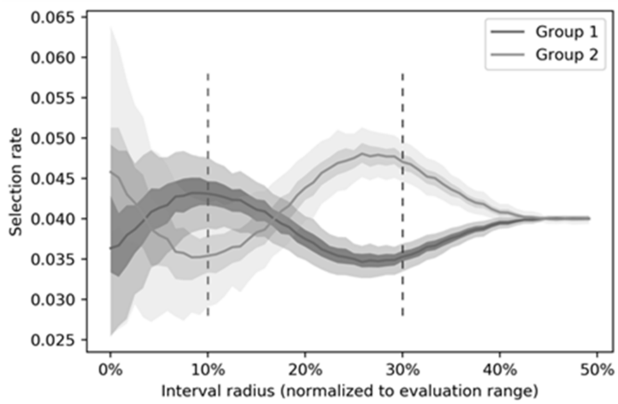
[202] *See* Schellman, *supra* note 147.

[203] In compliance with Title VII, which limits the use of quotas, accounting for uncertainties and biases should be considered an end, not a means to some target demographic makeup of the workforce.

[204] Johnson v. Transp. Agency, 480 U.S. 616, 630-31 (1987).

[205] Moto Kamiura & Kohei Sano, *Optimism in the Face of Uncertainty Supported by a Statistically-Designed Multi-Armed Bandit Algorithm*, 160 BIOSYSTEMS 25, 27 (2017).

Figure 7. Simulated selection rates for a threshold-based random lottery with varying interval lengths. As illustrated by the plot, using intervals of length 10% will result in a higher selection rate for Group 1 than Group 2, whereas using intervals of length 30% will result in a higher selection rate for Group 2. Note that interval length is a design choice, whereas errors and uncertainties are not; for example, the 10% interval length might correspond to a 75% confidence interval, whereas the 30% interval length might correspond to a 95% confidence interval.



By using data and mathematics to show that someone is qualified, the poset approach helps reduce the sense that any candidate is being interviewed but is not qualified for that interview. Of course, a strong cultural view that interviewing minority candidates is a compliance step will not be cured by posets alone. Management must support and explain the approach so that those involved in the process understand why it shows that candidates are indeed qualified. In addition, the approach must work within legal rules.

B. Posets and Legal Requirements

To understand how posets meet legal requirements, we start with what those requirements are. We then show how posets fit well within the requirements.

1. The Law's Approach to Numerical Assessments Supports Using the Poset Approach

Employers may use numerical assessments as part of hiring plans and assessments but there are boundaries on what they may do. If a plan is "blind hiring," that is, dictates hiring "solely by reference to statistics" or "by reflexive adherence to a numerical standard," the plan is not likely to be allowed.[206] But, if a plan takes "numerous factors . . . into account in making hiring decisions, including specifically the qualifications of [all] applicants for particular jobs," the plan may take a protected class into account as part of the overall evaluation.[207] In that sense, the protected class status "may be deemed a 'plus' in a particular applicant's file, yet it does not insulate the individual from comparison with all other candidates for the available seats."[208] Thus one is allowed to, and indeed must, compare candidates, but comparison does not require pure, numeric ranking. A pure numeric ranking might tip into the sort of "blind hiring" that is disfavored. Instead, discretion in comparison of candidates is allowed when it is part of the overall, individual assessment.

For example, in *Johnson v. Transportation Agency of Santa Clara County*, two candidates were deemed well-qualified based on a range of metrics, such as experience, background, and test scores taken together.[209] But each candidate had differences within a given metric. One had more clerical work and more road maintenance work; the other had more experience at a specific part of the business.[210] As for test scores, the employer had set 70 as the minimum threshold for the interview and the range of acceptable scores was 70 to 80.[211] Seven applicants scored above 70.[212] The male plaintiff scored 75 on the interview portion of the assessment and the promoted woman scored 73.[213] Thus, the woman was given the promotion over the man who had the higher score.[214] Because the scores were within the range of acceptable scores and the final hiring manager looked at a set of metrics with gender as "but one of numerous factors he took into account in arriving at his decision," the plan's incorporation of bias-awareness, here gender, was allowed.[215] *Johnson* is not the only case in which the concept of an absolute score or ranking is challenged.

[206] *Johnson*, 480 U.S. at 636-37.
[207] *Id.* at 637.
[208] *Id.* at 638.
[209] *Id.* at 616.
[210] *Id.* at 623.
[211] *Id.* at 623-24.
[212] *Id.*
[213] *Id.*
[214] *Id.*
[215] *Id.* at 638.

The act of "banding," or considering score ranges instead of singular scores, has been accepted to account for inaccuracies in evaluation.[216] In *Bradley v. City of Lynn*, the court invalidated the use of two civil service tests because (1) the industry standard was that validation studies should be conducted every five years, (2) to validate its test, the city relied on a 1992 report that was more than ten years old and that used even older data from a different jurisdiction, and (3) the tests were "not professionally-created nor professionally-validated."[217] In addition, the court noted that the city "adjusted scores after administering the examination by removing questions and by crediting multiple answers as correct on questions, so that the arbitrary passing point of seventy produced no adverse impact on minorities under the four-fifths rule."[218] These points explained why the test was not valid.

The *Bradley* court's critique shows how computer science can aid employment evaluation methods. Validation is required for "tests and other selection procedures which are used as a basis for any employment decision."[219] Employment decisions include "hiring, promotion, demotion," and, as discussed above, algorithms are used for such decisions.[220] Compared to relying only on expert reports and five-year studies, computer science methods are well-suited for demonstrating validity on a more frequent basis. Quality data was lacking in *Bradley*. Finding and analyzing better datasets is a fundamental part of good computer science. As for *ex post* adjustments, like with the defendants in *Ricci*, the *Bradley* defendants identified a problem and tried to fix it *ex post*. Not only is that a suspect, if not prohibited, practice; it may be avoidable. Recall that *Ricci* lauded efforts to design a test to reduce discriminatory outcomes. Computer science methods offer more tools to do just that. *Bradley* also identified the problem with rank ordering which recent computer science addresses.[221]

[216] Bos. Police Superior Officers Fed'n v. City of Boston, 147 F.3d 13, 24 (1st Cir. 1998) (recognizing a three-point band for test scores); Kirkland v. N.Y. State Dep't of Corr. Servs., 711 F.2d 1117, 1133 (2d Cir. 1983) (allowing a 4-point "zone" of test scores because "small differences between the scores of candidates indicate very little about the candidates' relative merit and fitness"); Bradley v. City of Lynn, 443 F. Supp. 2d 145, 173-74 (D. Mass. 2005).

[217] *Bradley*, 443 F. Supp. 2d at 172.

[218] *Id.*

[219] 29 C.F.R. § 1607.2 (B) (2022); *see* 29 C.F.R. § 1607.5 (B) (2022).

[220] *See supra* Figure 1; *supra* notes 89–91 and accompanying text.

[221] *Bradley*, 443 F. Supp. 2d at 173.

Even if relying on the 1992 report was allowed, the 1992 report validated rank ordering "only when a cognitive test constituted 40% and a physical test constituted 60% of the overall composite score."[222] The city used the written cognitive ability score "as the *sole* basis for ranking ordering."[223] The Sixth Circuit rejected that approach and said, "We reiterate that a selection procedure that ranks only on the basis of [cognitive ability test] scores is not acceptable."[224] Furthermore, the tests could not "be used reliably to distinguish candidates within a spread of as much as eight points."[225] As one expert put it, "there would be no difference between a score of 100 and a score of 92."[226] Another expert explained, "there is a spread . . . . Where [the score] doesn't matter."[227] In short, rank ordering exacerbates adverse impact.[228] In contrast, banding accounts for the fact that there may not be a "rational, statistically valid basis for distinguishing between candidates" and "diminishes" adverse impact in ways that are "consistent" with Title VII.[229]

Other courts have acknowledged the problems with test scores, pure ranking, and that sometimes race-conscious decisions are needed to remedy the problems that flow from testing and pure ranking. In *Kirkland v. N.Y. State Department of Correctional Services*, the Second Circuit has acknowledged "the fact that small differences between the scores of candidates indicate very little about the candidates' relative merit and fitness . . . ."[230] The Second Circuit has also held that one way to comply with Title VII is for an employer to "acknowledge his inability to justify rank-ordering and resort to random selection from within either the entire group that achieves a properly determined passing score, or some segment of the passing group shown to be appropriate."[231] The *Kirkland* court noted this rather broad approach is allowed while it endorsed a different one.

Under the other option, an employer assessed "a statistical computation of the likely error of measurement inherent" in its exam.[232] It then used that measurement to set up zones of candidates clustered by test scores

[222] *Id.* at 172.
[223] *Id.* at 173.
[224] Brunet v. City of Columbus, 58 F.3d 251, 255 (6th Cir. 1995).
[225] *Bradley*, 443 F. Supp. 2d at 173.
[226] *Id.*
[227] *Id.*
[228] *Id.*
[229] *Id.* at 174.
[230] Kirkland v. N.Y. State Dep't of Corr. Servs., 711 F.2d 1117, 1133 (2d Cir. 1983).
[231] *Id.* (citations omitted).
[232] *Id.*

within that error measurement.[233] That practice was seen as a good solution to "insur[e] compliance" with Title VII.[234] The Second Circuit explained, "by creating a more valid method to assess the significance of test scores, [the approach] eliminated the central cause of the adverse impact, i.e., the rank-ordering system, while *assuring appointments on the basis of merit*."[235] The First Circuit has also allowed employers to forego rank-ordering and choose a candidate whose score was "only one point shy of that of the candidates bypassed for his promotion" especially because expert testimony showed that "candidates who scored within a three-point band should be considered functionally equivalent . . . and equally qualified to successfully perform the job as any other person in that score band."[236] And recall that the Court in *Johnson* also recognized that a range of scores is allowed.[237]

In short, using the poset approach fits well within and is supported by case law. The cases support the use of multiple metrics to evaluate candidates, recognize there may be uncertainty in scoring, endorse applying methods to assess that uncertainty, and embrace using score ranges to compare candidates. The poset approach encompasses all these practices and offers stronger ways to show how each practice was conducted and valid.

2. Enabling Better Comparisons by Using Posets

Examining *Johnson* through our approach shows how the approach aids an entity in evaluating candidates. The Santa Clara Transportation Agency ("the Agency") reviewed its work force and found "while women constituted 36.4% of the area labor market, they composed only 22.4% of Agency employees."[238] The analysis went deeper and revealed women "were concentrated largely in EEOC job categories traditionally held by women."[239] Women were highly represented in the Office and Clerical Workers category but were "only 7.1% of Agency Officials and Administrators, 8.6% of Professionals, 9.7% of Technicians, and 22% of

233 *Id.*
234 *Id.*
235 *Id.* (emphasis added).
236 Boston Police Superior Officers Fed'n v. City of Boston, 147 F.3d 13, 24 (1st Cir. 1998).
237 Johnson v. Transp. Agency, 480 U.S. 616, 638 (1987).
238 *Id.* at 621.
239 *Id.*

Service and Maintenance Workers."[240] In addition, "for the job classification relevant to this case, none of the 238 Skilled Craft Worker positions was held by a woman."[241] Thus, the Agency faced exactly the sort of problem that using data can create. Statistically the company had far less data on the female candidate compared to the male candidate.

The Agency had this outcome with its approach: nine applicants "were deemed qualified for the job, and were interviewed by a two-person board."[242] That interview was scored and after that, seven of the nine were above the cutoff score which was 70.[243] The range of scores was 70 to 80. Johnson, the plaintiff, scored 75, which was second overall and tied another candidate.[244] Joyce, the woman who was promoted, "ranked next with a score of 73."[245] A second interview was conducted by three Agency supervisors, who ultimately recommended that Johnson be promoted."[246]

These interview scores, of course, are estimations and are not free from uncertainties. For example, one might compute confidence regions for the first round of interviews as depicted in Figure 7 (left). The second round of interviews might have resulted in slightly different scores and smaller confidence regions (see Figure 7, right), since having multiple evaluations might reduce uncertainty. At this point, one can use these confidence intervals (or uncertainty sets) to aid in making a selection between Joyce and Johnson. Based on the confidence regions in Figure 7, the two candidates are incomparable.[247] Johnson appears to have an edge in terms of interview scores, and Joyce has an edge on "other factors," so the candidates can be considered equally but differently qualified. Since the two candidates are *incomparable*, a decision can be made in several ways: (1) further resources can be invested in assessing the candidates, (2) the agency can manually weigh all the information, implicitly valuing certain factors over others, or (3) a decision can be made with a coin toss. Ultimately, the agency chose option (2), giving higher weight to the "other factors," including gender.

The poset approach better supports and explains why a decision was made especially when multiple factors are assessed. An issue in *Johnson*

[240] *Id.*

[241] *Id.*

[242] *Id.* at 623.

[243] *Id.*

[244] *Id.* at 623-24.

[245] *Id.* at 624.

[246] *Id.*

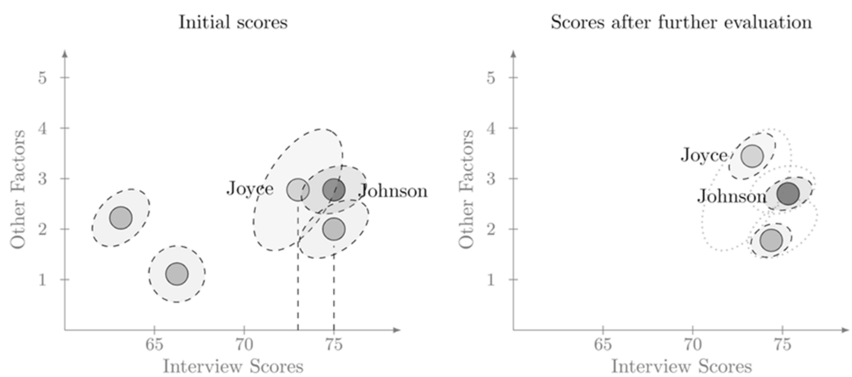
[247] Here, "incomparability" refers to the same notion described in Figure 4.

was whether the "Agency Plan unnecessarily trammeled the rights of male employees or created an absolute bar to their advancement."[248] In finding that the Plan did not, the Court noted that the Plan authorized consideration of a protected class, in this case gender, but that such consideration "was but one of numerous factors" used in making the final decision.[249] Recall that a plan may look at "race or ethnic background" as "a 'plus' in a particular applicant's file, yet it does not insulate the individual from comparison with all other candidates for the available seats."[250] The Agency Plan considered gender "but require[d] women to compete with all other qualified applicants. *No* persons are automatically excluded from consideration; *all* can have their qualifications weighed against those of other applicants."[251]

Thus, if the Agency has used the poset approach, it could have better demonstrated what factors were used, how they were assessed to create ranges of qualifying scores so that "their qualifications weighed against those of other applicants;" and show that candidates were not "automatically excluded from consideration."[252] Indeed, the Agency could have used the approach to show the assessments at each stage and further demonstrate the candidates were being evaluated against each other, not in isolation. Put simply, the poset approach provides a technical method for incorporating protected information into decision-making that does not reduce to blind hiring, and if two candidates are incomparable under the partial ranking, how protected class status was used as an allowed plus factor.

[248] *Johnson*, 480 U.S. at 637-38.
[249] *Id.* at 638.
[250] Regents of the Univ. of Cal. v. Bakke, 438 U.S. 265, 316-19 (1978).
[251] *Johnson*, 480 U.S. at 638.
[252] *Id.*

Figure 8. Possible confidence regions of candidates in *Johnson*. The left plot shows scores after the first round of interviews, and the right plot shows scores after the second round. The interview scores of Joyce and Johnson on the left are the candidates' actual scores, whereas the uncertainty sets (or confidence regions) are drawn for illustrative purposes.



C. Summary: A Framework for Using Posets in Practice

We end this Part by providing a framework for using the poset approach in practice. While this framework does not encompass every possible use of the poset approach,[253] it describes the process from beginning to end.

Step 1: Interrogate Your Employment Data. Recall that "unnecessary barriers to employment" must fall, even if "neutral on their face" and "neutral in terms of intent."[254] Furthermore hiring and promotion practices that "operate as 'built in headwinds' for minority groups"[255] are not allowed. These dictates beg a hiring entity to use data to identify the headwinds and be extension have evidence to support that an equal opportunity or DEI hiring plan is justified as a matter of law.

To see where problems may exist, a hiring entity will have to clean-up and process past hiring data. This task requires collection of data from previous hiring cycles. This data might include scores derived from textual analysis of resumes test scores for job-related tasks (e.g., computer programming test scores), automated scores based on analysis of video

[253] Salem & Gupta, *supra* note 154, at 9.
[254] Griggs v. Duke Power Co., 401 U.S. 424, 430-31 (1971).
[255] Ricci v. DeStefano, 557 U.S. 557, 622 (2009) (Ginsburg, J., dissenting).

interviews, college GPAs, courses taken, years of work experience, job performance of those who were hired, and so on.

Step 2. Quantify Uncertainty and Bias. Once the data is gathered, data analysis can be used to quantify potential data biases. Clusterings, for example, can help determine if evaluations unfairly favor one group over another. Looking at the data along different demographics (e.g., based on race, gender, age) can point to potentially discriminatory decisions in the past. Going further, one can use social science studies that highlight the impact of social status on the considered metrics (e.g., standardized test scores).[256] This process will help highlight qualitative and quantitative reasons for disparities in the past hiring data.

Step 3. Construct a Partial Order. Trends identified in Step 2 can be used to construct a partial ranking of candidates. For example, score ranges can be constructed for each attribute of interest using a prediction model and estimates of its error variances. These ranges can take into account distributional differences across protected attributes, differing error variances due to training data imbalance,[257] observed inaccuracies in past predictions, and so on. Unsupervised methods such as clustering can be used without the specific knowledge about protected information, or this can be abstracted out by a third-party vendor to simply provide a hiring entity with the resultant estimate of uncertainties or the poset over the candidates. The goal here is to account for uncertainties, inaccuracies, and biases in a direct and mathematically justified way, thereby paving the way to fairer decisions.

Step 4. Adapt Selection Algorithms. Once the partial ranking has been constructed, selections need to be made. Presumably, a hiring committee already has a screening process (automated or otherwise) which aligns with the goals of the employer. To implement the poset approach, this screening process must be adapted to take a partial ranking as input instead of numeric scores or a total ranking. Typically, this can be done by prioritizing maximality and randomizing wherever incomparabilities necessitate.[258]

---

256 *See, e.g.*, Steele & Aronson, *supra* note 109 (comparing participant performance, between Black and White participants, on 30 verbal items from the GRE with how the participants reported their thoughts on their academic competence and self-worth).

257 This refers to the observation that a group which is underrepresented in training data often experiences large errors in a resulting prediction model. In the poset approach, these larger errors could translate to larger score rangesfor the underrepresented group. Note that the groups in question could come from a clustering and need not be demographic groups.

258 Salem & Gupta, *supra* note 154, at 16 (providing an example in an online setting).

Step 5. Technical Accountability for Policy Compliance. The entire hiring pipeline may be subject to auditing for compliance with anti-discrimination policy. It is prudent to document and be able to justify each decision made in the hiring process, particularly those pertaining to the four steps outlined. Using these steps and maintaining records about data used, how algorithm design choices were made, uncertainties quantified, and so on, embraces the practices about design and data analysis the Supreme Court endorsed in *Ricci* and *Johnson*. Indeed, the nature of data analytics and using the mathematical approach inherent to posets should offer a stronger ability to explain why the process was sound should it face legal scrutiny.

Step 6. Return to Step 1 and Review and Update the Hiring Plan. A key factor in *Johnson* was that "the Plan sought annually to develop even more refined measures of the under-representation in each job category that required attention."[259] A data-driven plan, using posets or other methods, cannot set aside employment slots for a particular group.[260] As the Supreme Court has explained, a hiring entity may strive "to *attain* a balanced work force, not to maintain one."[261] This step helps ensure that a hiring entity is not using data to set up quota based on a particular imbalance at a given time.

This step requires that an entity assess whether the hiring plan has achieved its goals, and by extension showing what actions, if any, are needed at least annually. To be clear, an entity establishing a hiring plan and using a range of data-driven, bias-aware methods in that plan should also use the approach in Step 1 to answer whether the plan has reached its goals. If it has, the bias-aware methods may no longer be allowed under the law.

In a sense, the law's openness to creating a data-driven, bias-aware plan while also requiring that such a plan be assessed and recalibrated fits within an organizational and machine learning mindset.[262] The law is asking entities to use sophisticated methods to identify a problem, try a solution, assess outcomes, and update approaches on a continual basis. The steps set out here provide a map on how to do that.

[259] Johnson v. Transp. Agency, 480 U.S. 616, 635 (1987).
[260] *See id.* at 640-41.
[261] *Id.* at 639.
[262] Desai, *supra* note 27, at 568.

CONCLUSION

The desire to pursue DEI and the demands of current affirmative action law create problems for employers that computer science can help solve. Employers are always seeking and evaluating talent. The vast amount of candidate applications and historical data create the need to sort candidates at scale. Using algorithms to help solve this task is a common and obvious choice. Humans cannot easily sort thousands of applications by hand, and data offers the possibility of finding qualified candidates who hopefully will do well within an organization. Yet, known issues about the way datasets may entrench historical unfairness, combined with a larger question about the uncertainty inherent to a given evaluation metric, mean a company may unintentionally discriminate against a protected group. Not addressing such discrimination can lead to lawsuits, sanctions, and social condemnation. At the same time, announced efforts to increase minority hiring have faced legal scrutiny about whether companies are engaging in discrimination because they are taking race into account as part of their hiring plans. Furthermore, the law draws a distinction between affirmative action plans and the more general desire to increase diversity. This Article offers a way out of this morass.

By using current best practices in data science and operations research, companies can develop sound methods to pursue their diversity goals that pass legal muster. Any affirmative action plan must be ready for legal challenges. By extension, DEI plans will likely face challenges. A DEI plan that adheres to the rules for affirmative action plans should be well setup to survive legal challenge. In either case, an on-going question is whether a company may use algorithms in such a plan, and if so, what methods are allowed?

This Article has thus taken the step-by-step questions posed when pursuing affirmative action to show that the legal demands of each step embrace leveraging data science and algorithms to support such a plan. The law champions acting to remove "built-in headwinds" to employment opportunities.[263] Thus, we show how to use data science to document such barriers. The next question is what may be causing the barriers? We offer ways that computer science can identify biases in the selection process. We also review different computer science approaches to bias mitigation and show what they achieve, what they miss, and where they may devolve into quotas. To overcome these issues, we offer a new approach — the poset approach. The approach enables a company to acknowledge and

[263] Ricci v. DeStefano, 557 U.S. 557, 632 (2009).

account for uncertainty — actions that the law favors — and show that candidates are competing against "all other qualified applicants" so that "*all* are able to have their qualifications weighed against those of other applicants."[264] Furthermore, the approach enables a company to document the assumptions and mathematics behind its particular use of the approach. This ability means a company can show how it used multiple metrics to assess candidates and evaluate them as individuals as the law requires.

Put differently, rather than rejecting data science and algorithmic methods as leading to discrimination, interrogating legal rules to see what the law requires and what the law allows enables innovation on how to identify and mitigate bias.[265] Rigorous adherence to both legal requirements and technical methods means a company will be able to show evidence for each step of building its plan and especially that the plan was not a quota that automatically excluded a person or group. In short, this Article offers ways to use algorithms to tame discrimination as part of a path to algorithmic diversity, equity, and inclusion.

264 *Johnson*, 480 U.S. at 638.

265 To be clear, the methods here build tools and best practices, but those are only part of the solution. Management must put its full weight behind its efforts to pursue DEI so that the tools and practices are used and actual outcomes change.