LEGAL STUDIES RESEARCH PAPER SERIES

PAPER NO. 17-12-03

December 2017

Auditing Algorithms for Discrimination

166 University of Pennsylvania Law Review Online 189 (2017)

by

Pauline T. Kim
Daniel Noyes Kirby Professor of Law

Electronic copy available at: https://ssrn.com/abstract=3093982

ESSAY

AUDITING ALGORITHMS FOR DISCRIMINATION

PAULINE T. KIM†

I. INTRODUCTION

As reliance on algorithmic decisionmaking expands, concerns are growing about the potential for arbitrary, unfair, or discriminatory outcomes in areas such as employment, credit markets, and criminal justice. Legal scholars have lamented the lack of accountability of these automated decision processes and called for greater transparency.[1] They argue that the way to avoid unfair or discriminatory algorithms is to demand greater disclosure of how they operate.[2] *Accountable Algorithms* resists this call for transparency, calling it "a naive solution."[3] Instead, it argues that technology offers tools—"a new technological toolkit"—that can better assure accountability.[4]

One of the examples that Kroll et al. rely on to illustrate their argument is the goal of ensuring that algorithms do not discriminate. Many

† Daniel Noyes Kirby Professor of Law, Washington University School of Law. Thanks to my colleagues here at Washington University, especially Peggie Smith and Neil Richards, and to anonymous reviewers for the FAT* Conference, for their helpful comments.

1 *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 1-18 (2015) (raising concerns about secret algorithms used by internet and finance companies and calling for a more "intelligible" society which entails, among other things, greater transparency); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 8 (2014) (criticizing the secrecy surrounding credit scoring systems and arguing that "transparency . . . is essential"); Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 421 (2014) ("Transparency has heightened importance with the arrival of big data.").

2 *See, e.g.*, Citron & Pasquale, *supra* note 1, at 24-25 (arguing that the FTC should be given access to scoring systems to check for bias and unfairness); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 126 (proposing that those who use data for predictive purposes disclose the data used and the methodology employed).

3 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, 657 (2017).

4 *Id.* at 633.

Electronic copy available at: https://ssrn.com/abstract=3093982

commentators have pointed out the risk that automated decision processes may produce biased outcomes,[5] and in prior work, I have argued that serious policy concerns are raised when these algorithms exacerbate historic inequality or disadvantage along the lines of race, sex, or other protected characteristics—what I've referred to as "classification bias."[6] Recognizing that the precise meaning of discrimination is uncertain and contested, Kroll et al. do not try to resolve debates over the meaning of discrimination.[7] Instead, without choosing among the competing definitions, they simply survey the available technical tools, suggesting that these tools will be more effective at ensuring nondiscrimination than calls for transparency.

Transparency involves outside scrutiny of a decision process, for example, by allowing third parties to examine the computer code or the decision criteria it implements. Auditing is another method for promoting transparency.[8] When the goal is nondiscrimination, auditing could involve techniques to ensure that an algorithm follows a specified rule—for example, sorting must not occur based on race or sex. Alternatively, auditing for discrimination could take the form of examining inputs and outputs to detect when a decision process systematically disadvantages particular groups. The latter form of auditing does not involve direct examination of the decision process, but is useful in detecting patterns. This type of auditing, in the form of field experiments, is well established in the social science literature as a technique for testing for discrimination in decisions such as employment and consumer transactions.[9] Auditing the effects of decisionmaking algorithms

5 *See, e.g.*, EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 51-53 (2014); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677-93 (2016); Citron & Pasquale, *supra* note 1, at 4; Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, And It's Not Fair*, 66 STAN. L. REV. ONLINE 35, 36-37 (2013).

6 *See* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 890-91 (2017)

7 Kroll, et al., *supra* note 3 at 678-79.

8 Kroll et al. define auditing as techniques used to independently evaluate whether computer systems conform "to applicable regulations, standards, guidelines, plans, specifications, and procedures." Kroll, *supra* note 3, at 660-61 (quoting IEEE Computer Society, IEEE Std 1028 - IEEE Standard for Software Reviews and Audits § 8.1 (Aug. 15, 2008), http://ieeexplore.ieee.org/document/4601584 [https://perma.cc/WLD6-VPUN]). The authors are not entirely clear whether they consider auditing to be another form of transparency or something distinct. *See, e.g.*, id., *supra* note 3, at 660 ("Beyond transparency, auditing is another strategy for verifying how a computer system works."). Auditing may not entail complete transparency regarding the underlying code or precisely reveal how a program makes decisions. Nevertheless, because auditing techniques reveal significant information about how computer systems operate, I believe they are appropriately characterized as a method that promotes transparency.

9 *See, e.g.*, Ian Ayres, *Fair Driving: Gender and Race Discrimination in Retail Car Negotiations*, 104 HARV. L. REV. 817 (1991); 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 (2004); Joanna N. Lahey, *Age, Women, and Hiring: An Experimental Study*, 43 J. HUM. RESOURCES 30 (2008); David Neumark et al., *Sex Discrimination in Restaurant Hiring: An Audit Study*, 111 Q. J. ECON. 915 (1996); David Neumark, *Detecting Discrimination in Audit and*

similarly offers a method of detecting when they may be biased against particular groups. Kroll et al., however, express skepticism about auditing as a strategy, arguing that it is not only technically limited, but also likely restricted by law.[10] More specifically, they suggest that when an algorithm is found to have a disparate impact, the Supreme Court's decision in *Ricci v. DeStefano*[11] may prevent correcting for that bias.[12]

This Essay responds to Kroll et al., arguing that, despite its limitations, auditing for discrimination should remain an important part of the strategy for detecting and responding to biased algorithms. Technical tools alone cannot reliably prevent discriminatory outcomes because the causes of bias often lie not in the code, but in broader social processes. Therefore, implementing the best available technical tools can never guarantee that algorithms are unbiased. Avoiding discriminatory outcomes will require awareness of the actual impact of automated decision processes, namely, through auditing.

Fortunately, the law permits the use of auditing to detect and correct for discriminatory bias. To the extent that Kroll et al. suggest otherwise, their conclusion rests on a misreading of the Supreme Court's decision in *Ricci*. That case narrowly addressed a situation in which an employer took an adverse action against identifiable individuals based on race, while still permitting the revision of algorithms prospectively to remove bias. Such an approach is entirely consistent with the law's clear preference for voluntary efforts to comply with nondiscrimination goals.

II. THE LIMITS OF A TECHNOLOGICAL RESPONSE

Kroll et al. resist the call for transparency, arguing that it may jeopardize other important interests and will often be ineffective in ensuring fairness. They advocate instead for reliance on technological tools to ensure accountability. Although greater understanding of these tools will aid policymakers, this Part explains that technological fixes alone will not be able to prevent discriminatory outcomes because the causes of bias often lie outside the computer code.

According to Kroll et al., there are both policy and technical reasons to avoid relying on transparency as a solution. First, transparency is often in tension with other important interests, such as protecting trade secrets, ensuring the privacy of sensitive personal information, and preventing

Correspondence Studies, 47 J. HUM. RESOURCES 1128 (2011) (describing the process of an audit study, namely the use of fictitious individuals kept identical besides the trait being studied).

10 *See* Kroll et al., *supra* note 3, at 694 ("The holding in *Ricci* suggests that we cannot solely rely on auditing for legal reasons.")

11 557 U.S. 557 (2009).

12 *See* Kroll et al., *supra* note 3, at 694-95.

strategic gaming of automated decision systems.[13] Even when transparency does not compromise other interests, it may not be effective as a practical matter. As Kroll et al. explain, merely looking at the source code may not reveal what a program does.[14] Dynamic testing will better show how a program actually operates, but because testing all possible inputs is impracticable, there is no assurance that a model will operate fairly across all cases.[15] Dynamic machine learning techniques create further challenges, because the decision process itself is constantly changing.[16] In short, the limitations of computer science mean that "for any nontrivial property of a program's behavior," there are no technical tools that can guarantee that a particular program always has that property.[17]

This is all sobering news for those who have pinned their hopes for algorithmic fairness on greater transparency. But Kroll et al. suggest a way out of the transparency trap. "Fortunately," they write, "technology is creating new opportunities—more subtle and flexible than total transparency—to make automated decisionmaking more accountable to legal and policy objectives."[18] They describe several technological tools that make it possible to confirm that a decision process has certain properties demanded by fairness, even though the actual decision criteria may remain secret.

One of the examples Kroll et al. use to illustrate their arguments is the goal of ensuring nondiscrimination.[19] Designing a system to be accountable for a substantive goal like nondiscrimination is difficult because it requires specifying the policy goals in terms precise enough to be reduced to code.[20] What constitutes forbidden discrimination is highly contested in the legal and political spheres, and these debates pose a problem for computer

13 *See id.* at 639, 658.

14 The lines of code comprising a program are often "complicated or obfuscated," and reading them tells little about how the program operates in the real world. *Id.* at 647.

15 *See id.* at 650-51.

16 This means that transparency at one moment cannot explain the algorithm's decision process at a different point in time. *See id.* at 638, 659-60.

17 *Id.* at 652.

18 *Id.* at 640.

19 Kroll et al. discuss two types of challenges. The first involves procedural regularity, namely, assuring that a decision process is applied consistently across cases. *See id.* at 656-77. The technological toolkit works better for this type of problem than for advancing substantive goals such as nondiscrimination. *See id.* at 678 ("Fidelity to policy choices like nondiscrimination is a more complicated goal than procedural regularity, and the solutions that currently exist to address it are less robust."). If accountability is part of the initial design, it is possible to build a system that guarantees procedural regularity, and permits third parties to verify that the chosen decision policy has been consistently applied.

20 As Knoll et al. put it, the challenge is to "bridg[e] the gap between technologists' desire for specificity and the policy process's need for ambiguity." *Id.* at 642. *See also id.* at 678 (stating that computer scientists "generally require a well-defined notion of what sort of fairness they are supposed to be enforcing.").

programmers. Without trying to resolve disagreements about the meaning of nondiscrimination, Kroll et al. describe the available techniques that computer scientists have developed. Each tool is designed to avoid certain outcomes that might be viewed as objectionable, depending upon one's definition of discrimination.

The authors' discussion of the available technical tools is important for policymakers to understand. Without repeating the technical details, several crucial points emerge. First, a simple prohibition on the use of protected characteristics such as race and sex in an automated decision process is easy to implement, but would do little to prevent biased outcomes.[21] In any sufficiently rich dataset, proxy variables likely exist that closely correlate with these characteristics, permitting implicit sorting on those bases. Second, some of the technical strategies for nondiscrimination require awareness and use of protected characteristics in order to ensure fairness across groups. Thus, enforcing notions of nondiscrimination beyond naïve "blindness" will often require decisionmaking systems to collect and utilize information about protected characteristics.[22] And finally, any technical response to the problem of discrimination requires thinking about the issue from the outset.[23] Because post hoc requirements of accountability may be difficult to enforce technically, systems must be built in ways that avoid discrimination, or at the very least, allow for accountability after the fact.

These insights are crucial for moving toward fair and accountable algorithms, but Kroll et al.'s survey of existing tools suggests we are a long way from a technical fix when it comes to preventing discrimination. As they recognize, one challenge is the ongoing disagreement about the meaning of discrimination. This lack of consensus is compounded by the fact that it may not be possible to satisfy different definitions of fairness simultaneously. Consider, for example, a system that seeks to accurately predict outcomes, such as parole violations, in a manner that is fair and equal across two groups. Nondiscrimination might be defined as equalizing the proportion of correct negative and correct positive predictions for each group. Alternatively, it might be defined as equalizing the proportion of false positives or equalizing the proportion of false negatives across the groups. It turns out that where base rates differ between the groups—e.g., if a higher proportion of men violate parole than women—then these three notions of fairness will be

21 *See id.* at 685 ("Blindness to a sensitive attribute has long been recognized as an insufficient approach to making a process fair.").

22 *See id.* at 687 (describing a process of "fairness through awareness"); *see, e.g.*, Cynthia Dwork et al., *Fairness Through Awareness*, 2012 PROC. 3RD INNOVATIONS THEORETICAL COMPUTER SCI. CONF. 214; *see also* Kim, *supra* note 6, at 918 (arguing that the risk of omitted variable bias means that controlling for sensitive demographic variables may sometimes be necessary to avoid biased results).

23 *See* Kroll et al., *supra* note 3, at 640.

incompatible, such that satisfying one of them will necessarily mean that both of the other criteria cannot be met.[24]

The problem, however, is not solely one of choosing among competing definitions of discrimination. Even if a clear consensus existed, there is a more fundamental reason technical tools alone cannot solve the problem. Namely, these tools do not and cannot address many of the reasons that automated decision processes may be biased in the first place. Existing scholarship has catalogued many potential sources of algorithmic bias—for example, biased data inputs, skewed training data, missing variables, selection of biased target variables, measurement errors, or intentional efforts to mask discriminatory motives.[25] Because algorithms may be biased for many different reasons, a technical response will only be successful to the extent that the available tools address the sources of bias in a given decision process. To see the limitations of a purely technical approach, consider two of the tools offered to address biased algorithms.

One of the technical strategies Kroll et al. discuss is introducing randomness into a decision process to allow it to recognize and overcome hidden biases.[26] They illustrate this strategy by using the example of a hiring algorithm trained using biased data that erroneously labels women as weak candidates. If the model mostly recommends men, it would "create a self-fulfilling prophecy" because the characteristics of successful hires will strongly correlate with gender.[27] To counter this tendency, they suggest that the algorithm "incorporate an element of randomness such that some candidates who are not predicted to do well get hired."[28] The subsequent performance of these random hires can then be observed, and if they perform better than predicted, the model can be updated, improving the fairness of the system over time.

While incorporating randomness may be useful in correcting some types of erroneous predictions—such as which online ads are more likely to generate clicks—it is far less likely to be successful in eliminating bias in other contexts, such as predicting job performance, credit worthiness or recidivism

24 *See* Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns & Aaron Roth, *Fairness in Criminal Justice Risk Assessments: The State of the Art*, arXiv:1703.09207v2 [stat.ML] May 28, 2017; Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, arXiv:1609.05807v2 [cs.LG] Nov 17, 2016.

25 *See, e.g.*, Barocas & Selbst, *supra* note 5, at 677-93 (2016) ("This Part develops a taxonomy that isolates and explicates the specific technical issues that can give rise to models whose use in decision making may have a disproportionately adverse impact on protected classes."); Kim, *supra* note 6, at 878-83 (identifying additional reasons that predictions based on atheoretical data mining techniques may produce biased predictions).

26 *See* Kroll et al., *supra* note 3, at 683-84.

27 *Id.* at 684.

28 *Id.*

risk. This is because incorporating randomness will not always address the underlying reasons the model is biased in the first place. To expand on Kroll et al.'s example, a hiring algorithm might erroneously classify women candidates as poor prospects because the variable chosen to measure successful job performance is inapt (e.g., measuring physical strength when manual dexterity contributes more to productivity), or incorporates biased judgments (e.g., subjective evaluations by biased supervisors). Alternatively, women may actually be less productive in certain fields because workplace structures limit their access to important mentoring or professional growth opportunities, or because a culture of pervasive sexual harassment hinders their work performance. In any of these situations, including more cases through randomization without addressing the underlying source of the problem will not debias the model and may even reinforce its discriminatory effects. In other words, when a model produces biased outcomes due to the processes generating the input values, merely tweaking the distribution of data inputs will not solve the problem.

A strategy based on randomization also faces significant practical limitations when applied in many real world contexts. Judges or parole boards would have to be willing to release some randomly selected criminal defendants; banks and employers similarly would have to be willing to extend credit or offer jobs to some randomly chosen applicants. And in any situation in which the outcome turned on human judgment, the method of selection in specific cases would have to remain hidden. For example, an employer should not know which of its actual hires was randomly selected, nor should it be able to infer that members of certain subgroups were more likely to have been randomly selected. Without complete blinding to the selection process, any subsequent assessments of job performance might be biased.

In addition, the effectiveness of randomization depends in part on the distribution of attributes within the pool and the proportion of cases selected randomly. In the hiring context, if the position is highly selective and the qualifications of applicants varies widely, a very high proportion of random hires will need to be observed before any bias can be detected. And if most applicants are unqualified, randomly selected candidates are likely to perform poorly, such that observing those outcomes may have the effect of reinforcing the existing bias.

Kroll et al. discuss another strategy, drawn from the work of Dwork et al., which seeks to constrain outcomes in a way that ensures nondiscrimination.[29] Dwork et al. formalize the idea of fairness in terms of "enforc[ing] similar probabilities of each possible outcome on similar people."[30] In other words,

29 *Id.* at 687.

30 *Id.*

the difference in the probability of outcomes for two individuals should be less than any differences between them. This strategy "requires a mathematically precise notion of how 'different' people are" and "must also capture all relevant features" in order to be implemented.[31] If these conditions can be met, computer scientists can build systems that satisfy the specified fairness constraint. One challenge, of course, lies in identifying "all relevant features," an obviously value-laden exercise.

However, the difficulty is only partly political. Even where there is consensus on which features matter, measuring them in an unbiased way may pose challenges. Consider again the hiring context. There may be agreement that leadership or an ability to work effectively on a team is relevant to a particular job, but an automated decision system that satisfies the fairness constraint—that similar people have similar probabilities of each of the outcomes—is not guaranteed to be unbiased. The system may have discriminatory effects if "leadership" and "collegiality" are measured in ways that reflect biased human judgments or workplace structures that systematically disadvantage certain groups. Thus, technical tools alone will not fix discriminatory algorithms where the bias is rooted in social problems, not purely technological ones.

Because so many potential sources of bias lie outside the code, no amount of technical design can ensure that automated decision systems will never operate in a discriminatory manner. Even the most carefully designed systems may inadvertently encode preexisting prejudices or reflect structural bias. For this reason, avoiding discrimination requires not only attention to fairness in design, but also scrutiny of how these systems operate in practice. Only by observing actual outcomes is it possible to determine whether there are discriminatory effects. In other words, if nondiscrimination is a substantive goal, auditing for the actual racial or gender impacts of automated decision processes will be necessary to diagnose when bias might have crept in and influenced the process.

Kroll et al. are skeptical of the utility of auditing. They assert that merely looking at inputs and outputs is limited, because it "tells an analyst very little about *why* differential behavior was observed."[32] They are right that auditing alone cannot precisely identify when and how impermissible discrimination occurs. Nevertheless, because the technical tools cannot ensure that algorithms are unbiased, auditing must remain an essential part of the toolkit for combating discrimination.

---

31 *Id.*

32 *Id.* at 651.

III. AUDITING AFTER *RICCI*

Auditing the actual outcomes produced by an algorithm can reveal when it disproportionately screens out protected groups, allowing the user to engage in self-examination and to revise its processes to eliminate implicit or unintended biases. However, Kroll et al. suggest that using auditing in this way may not be possible for legal reasons. They argue that under *Ricci v. DeStefano*, "users of algorithms may be legally barred from revising processes to correct for discrimination after the fact."[33] This claim stems from a misreading of the Supreme Court's decision. In fact, nothing in *Ricci* prohibits revising an algorithm after discovering that it has discriminatory effects. Doing so is not only legally permitted, but is precisely the type of compliance effort that the law encourages.

Ricci was decided under Title VII,[34] the statute which prohibits employers from discriminating on the basis of race, color, religion, national origin, or sex. Because *Ricci* involved Title VII's application to a workplace dispute, the discussion here also focuses on employment discrimination law, although the legal principles discussed are likely generalizable to other types of discrimination law. The Supreme Court found liability in *Ricci* when a public employer discarded the results of a promotional exam based on the racial profile of the successful test takers.[35] However, as explained below, the facts in *Ricci* are entirely distinct from situations in which a decisionmaker, like an employer, seeks to change an algorithm prospectively to remove bias.

Imagine that an employer receives thousands of applications each month for relatively low-skill positions, such as its call center operators. An automated tool collects each applicant's responses to an online questionnaire, aggregates them with other personal information available from private data brokers, and applies a sorting algorithm to select the top 10% of candidates for further review. After relying on this process for several months, the employer becomes aware that the decision process disproportionately screens out applicants from racial minority groups. Concerned about the biased outcomes, the employer reexamines the decisionmaking protocol, recognizes that some of the data inputs are unjustifiably biasing the results, and

33 *Id.* at 692. Barocas and Selbst similarly assert that "[a]fter an employer begins to use the model to make hiring decisions, only a 'strong basis in evidence' that the employer will be successfully sued for disparate impact will permit corrective action." Barocas & Selbst, *supra* note 5, at 726 (quoting Ricci v. DeStefano, 557 U.S. 557, 585 (2009)).

34 Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(a) (2012). Other laws protect against discrimination in employment on other bases (e.g., age, disability, genetic characteristics) or in different settings (e.g., housing, education, credit markets), and these laws differ from Title VII in certain ways. Although discussion of these other laws is beyond the scope of this Essay, basic concepts about the nature of discrimination are similar across statutes.

35 *Ricci*, 557 U.S. at 579-80.

implements a change to make the decision process fairer. Nothing in *Ricci* legally prohibits the employer from doing so.

Ricci addressed a challenge to the New Haven Fire Department's promotion process.[36] The City of New Haven had engaged in an extensive process to develop promotional exams for lieutenant and captain positions.[37] The City then announced the format of the tests, identified relevant study materials, and set a three-month preparation period.[38] After the exams were administered and scored, it became clear that they had a significant racial impact and, under the City's civil service rules, virtually all of the promotions would go to white firefighters, even though a significant proportion of the candidates were black or Hispanic.[39] Concerned that it would be vulnerable to a disparate impact suit if it made promotions on that basis, the City declined to certify the results of the exams.[40] Frank Ricci and others filed suit, alleging that they would have been promoted if the results had been certified and that the City's failure to do so constituted race discrimination.[41]

A majority of the Supreme Court agreed with the plaintiffs. Important to understanding that outcome is recognizing its starting point. The majority opinion states: "Our analysis begins with this premise: The City's actions would violate the disparate-treatment prohibition of Title VII absent some valid defense."[42] It found that "[t]he City rejected the test results solely because the higher scoring candidates were white,"[43] and thus, had taken an adverse action because of race—an action which would violate Title VII unless it had a valid defense.[44] Given this starting point, the majority opinion focused on whether to excuse the City's action because the City believed it necessary to avoid disparate impact liability. The majority held that actions like the City's violate Title VII unless the employer can demonstrate "a strong basis in evidence" that it would otherwise be liable for disparate impact discrimination.[45] After an examination of the facts, the Court concluded that the City lacked the necessary "strong basis in evidence" and found a violation of Title VII.

Notice that the "strong basis in evidence" requirement only becomes relevant if the employer has engaged in intentional discrimination and seeks to defend its actions as necessary to avoid disparate impact liability. In *Ricci*, the disparate treatment violation occurred *not* because the City expressed

36 *Id.* at 563-75.
37 *Id.* at 563-65.
38 *Id.* at 565.
39 *Id.* at 566.
40 *Id.* at 566-74.
41 *Id.* at 562-63.
42 *Id.* at 579.
43 *Id.* at 580.
44 *Id.* at 579.
45 *Id.* at 563.

concern about racial fairness, but because its rejection of the test results "adversely affected specific and visible innocent parties."[46] As Justice Kennedy wrote, "The injury arises in part from the high, and justified, expectations of the candidates who had participated in the testing process on the terms the city had established."[47] By refusing to certify the results, the City upset the "legitimate expectations" of those who took the test, including many who "invested substantial time, money, and personal commitment in preparing for the tests."[48] Similarly, Justice Alito's concurrence emphasized the plaintiffs' reliance on the City's announced procedures. For example, Frank Ricci, who has dyslexia, had hired someone to audio record the study materials, and Benjamin Vargas gave up a part-time job and his wife took leave from her job to enable him to study for the exam.[49]

Most workplaces are quite different from New Haven's Fire Department, which faced considerable constraints in developing its promotion process. Under civil service rules and its contract with the firefighters' union, the City was required to utilize an examination process that included written and oral components, weighted 60% and 40% respectively, and to provide detailed information about the exams well in advance of administering them.[50] By contrast, employers in the private sector have a great deal of discretion in determining hiring criteria and changing them as they see fit, so long as they do not discriminate. Thus, unlike the plaintiffs who sued in *Ricci*, applicants to our hypothetical company described above have not suffered an adverse action because of their race merely because the employer decided to change its hiring algorithm.[51] Applicants would have no legitimate expectations that the company's hiring criteria would never change, and could not credibly claim to have acted in reliance on a particular version of a complex and opaque algorithm. As a result, if the employer chose to revise the algorithm to eliminate unintended biases, no legitimate expectations would be disrupted and nothing in *Ricci* would prevent the employer from making the change prospectively. Because no disparate treatment violation would have occurred, the "strong basis in evidence" standard is simply not relevant.

---

46 Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1362 (2010).

47 *Ricci*, 557 U.S. at 593.

48 *Id.* at 583-84.

49 *See id.* at 607 (Alito, J., concurring).

50 *Ricci*, 557 U.S. at 564-65.

51 Applicants might claim that they have suffered an adverse action because their odds of being selected have been reduced by the change in the hiring algorithm. This claim, however, assumes the fairness of the prior distribution of the odds of success. It may be that the background conditions that influenced the initial distribution of probabilities were themselves unfair, such that those who benefited from them are not entitled to what amounts to an unfair advantage. For an exploration of these issues in the context of the affirmative action debate, see Pauline T. Kim, *The Colorblind Lottery*, 72 FORDHAM L. REV. 9 (2003).

The fact that the employer's motive for changing an algorithm is to avoid a racially biased outcome does not make it a violation of Title VII. All of the justices agreed on this point in *Ricci*. Justice Kennedy wrote for the majority: "Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race."[52] It follows that an employer is also permitted to consider how to redesign a test or practice so that it is fair to all in the future. Similarly, Justice Ginsburg, writing for the four dissenting justices, asserted that the fact that city officials were "conscious of race during their decisionmaking process . . . did not mean they had engaged in racially disparate treatment."[53] Mere race consciousness does not constitute disparate treatment if no adverse action has been taken against particular individuals.

After *Ricci*, then, employers are permitted to audit automated decision processes and change them prospectively in order to eliminate identified biases. Cases applying *Ricci* in the lower courts confirm this conclusion. For example, in *Maraschiello v. City of Buffalo Police Department*, the Second Circuit rejected the claim of a white officer that the City had discriminated against him when it adopted a new promotion exam.[54] Because part of the City's reason for revising the test was to avoid litigation challenges by minority officers,[55] Maraschiello argued that the City's decision to use the new test violated Title VII under *Ricci*. The court rejected his argument, concluding that no disparate treatment occurred.[56] Unlike in *Ricci*, where test results had been discarded because of their racial makeup, in *Maraschiello*, the City simply carried through on "a long-planned adoption of a new standard."[57]

---

52 *Ricci*, 557 U.S. at 585.

53 *Id.* at 619 (Ginsburg, J., dissenting).

54 709 F.3d 87, 89 (2d Cir. 2013). Maraschiello had received the highest score on a prior version of the exam administered in 2006. *Id.* No open positions were available for most of the period that the eligibility list was in effect. *Id.* In 2008, the City administered a revised promotional exam. During the period the revised test was being administered, a position opened up. The City allowed the original list to expire and then chose a candidate for the open position from the eligibility list determined by the revised test. *Id.* at 90-91.

55 The City was motivated in part to review the old test because it was aware of civil rights lawsuits challenging similar exams; however, it also acted in response to an expert's conclusion that the exam, which was based on a 30-year-old job analysis, was not a valid test of current job requirements. *Id.* at 89.

56 It did not help Maraschiello's case that he chose not to take the revised exam, and that the position was eventually given to another white male. *Id.* at 90-91.

57 *Id.* at 96. The Court explained: "Unlike in *Ricci*, where the results of a specific test were simply discarded based on the racial statistics reflected in the results, here the City replaced the 2006 list with the 2008 list after spending more than a year preparing to revise its assessment methods. Its problem was with the test itself, rather than with a particular set of results . . . .

Completing the last phase of a long-planned adoption of a new standard is a far cry from rejecting a set of results out of hand because of their racial makeup. Updating an examination . . . does not 'create[]

The court found that "[e]ven if it were determined that the City's choice to adopt a new test was motivated in part by its desire to achieve more racially balanced results," its actions did not amount to "race-based adverse action."[58]

Similarly, in *Carroll v. City of Mount Vernon*, the court rejected the claim of a white firefighter who alleged discrimination because the City allowed an eligibility list, on which he ranked near the top, to expire.[59] The City had delayed filling an open position in order to allow its legal department to review complaints from minority firefighters about the use of the earlier list.[60] While the legal review was pending, the list expired and a new eligibility list based on a more recent exam became effective.[61] The court found no evidence of discriminatory intent.[62] Far from the "express, race-based" decision in *Ricci*, Carroll's promotional opportunity "simply evaporated" while the City sought legal advice.[63] As the court explained, "Employers should be permitted to 'self-evaluate,' and take that 'hard look,' without fear that doing so would itself violate Title VII."[64]

In *Maraschiello* and *Carroll*, the plaintiffs' claims failed because their employers' decisions to change selection processes prospectively did not constitute adverse actions under Title VII. And in the absence of a disparate treatment violation, the employers did not need to meet *Ricci*'s "strong basis in evidence" standard to justify their revised tests. The results in these cases—permitting employers to change prospectively their selection procedures in order to reduce bias—are consistent with *Ricci*[65] and follow

---

a materially significant disadvantage with respect to the terms of . . . employment.'" *Id.* at 95-96 (quoting Williams v. R.H. Donnelly Corp., 368 F.3d 123, 128 (2d Cir. 2004)).

58 *Id.* at 95-96.

59 707 F. Supp.2d 449 (S.D.N.Y. 2010), *aff'd by* Carroll v. City of Mount Vernon, 453 F. App'x 99 (2d. Cir. 2011).

60 *Id.* at 452.

61 *Id.* Carroll ranked sixteenth and was therefore listed too low to be eligible for the next available positions.

62 *Id.* at 455.

63 *Id.* at 456.

64 *Id.* at 458 (quoting Moody v. Moody, 422 U.S. 405, 417-18 (1975), and Ricci v. DeStefano, 557 U.S. 557, 587 (2009)). The court also stated that, "If the Court were to find that defendants cannot take time to consider potentially valid objections merely because they relate to the representation of racial minorities, it would not only discourage compliance with court-sanctioned efforts at achieving diversity and equal opportunity . . . but it would also discourage the voluntary compliance that is 'the preferred means of achieving the objectives of Title VII.'" *Id.* at 457-58 (quoting Int'l Ass'n of Firefighters v. City of Cleveland, 478 U.S. 501 (1986)).

65 As I explain elsewhere:

> The majority in *Ricci* objected to *undoing* the results of the test once the employer announced and administered it; the Court did not require the City to continue using the test results to make future promotion decisions. To suggest otherwise would lead to the absurd result that an employer, who ordinarily has a great deal of discretion to change its selection processes or criteria, would suddenly be prohibited from changing a practice the moment it learned that it had a disparate effect on a protected group. Such an outcome would produce the exact *opposite* effect that Congress

directly from the Supreme Court's longstanding encouragement of voluntary efforts to eliminate employment discrimination.[66] Similarly, if the user of an algorithm believes that it is causing unjustified bias, the law permits correction of the process going forward.

For Kroll et al., the concern that *Ricci* might prohibit "correct[ing] for discrimination after the fact" bolsters their call to design for fairness "from the start."[67] They argue that "incorporating nondiscrimination in the initial design of algorithms is the safest path that decisionmakers can take."[68] Undoubtedly, designing algorithms to be nondiscriminatory is by far the preferable practice. However, as explained in Part II above, even the best technical tools cannot guarantee that algorithms will not discriminate because bias may result from social processes that lie outside the code. Thus, it is crucial not to rely too heavily on technical tools implemented ex ante, or to foreclose the possibility of making corrections upon discovering discriminatory effects after an algorithm is in use.

IV. Conclusion

As computer scientists develop better technological tools to ensure that algorithms are fair, calls for transparency may appear unnecessary or even counterproductive. When it comes to the goals of nondiscrimination, however, purely technical strategies cannot guarantee that automated decision processes will be free of bias. These processes may systematically disadvantage protected groups as a result of social processes that lie outside the code, and therefore, fixes internal to a computer program will not be sufficient. For this reason, auditing—examining the actual impact of algorithms on protected classes—should remain an important part of the toolkit. Auditing is an essential strategy for detecting unintended bias and prompting the reexamination and revision of algorithms to reduce discriminatory effects. Fortunately, auditing and correcting for bias is not

intended Title VII to have—namely, it would freeze into place employer practices that work to systematically disadvantage minority applicants and employees. The way to avoid such an absurd result is to recognize that acting prospectively to prevent classification bias is not a form of intentional discrimination.

Kim, *supra* note 6, at 932 (internal citation omitted).

66 As the Court wrote in *Int'l Ass'n of Firefighters v. City of Cleveland*, "We have on numerous occasions recognized that Congress intended voluntary compliance to be the preferred means of achieving the objectives of Title VII." 478 U.S. 501, 515 (1986). Elsewhere, it explained that Title VII was intended to act as a "spur or catalyst" prompting employers "to self-examine and to self-evaluate their employment practices." Albemarle Paper Co. v. Moody, 422 U.S. 405, 417-18 (1975) (quoting United States v. N. L. Indus., Inc., 479 F.2d 354, 379 (8th Cir. 1973)).

67 Kroll et al., *supra* note 3, at 640, 692.

68 *Id.* at 695.

only legally permissible, it also represents the type of voluntary compliance effort that Supreme Court precedents have long endorsed.

Preferred Citation: Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189 (2017), http://www.pennlawreview.com/online/166-U-Pa-L-Rev-Online-189.pdf.