NORTH CAROLINA JOURNAL of
LAW & TECHNOLOGY
Volume 28, Issue 1 2026

ARTICLE

COLORBLIND ALGORITHMS AFTER *SFFA*: RACE-CONSCIOUS COMPLIANCE AND THE FUTURE OF CIVIL RIGHTS IN AI GOVERNANCE

Susan Tanner†

Artificial intelligence can reproduce racial inequality without ever using race as an explicit input. Hiring models, tenant-screening systems, lending tools, and other automated decision systems can encode racialized patterns through proxies, historical data, and seemingly neutral concepts such as merit, fit, risk, and creditworthiness. Yet the legal response to Students for Fair Admissions v. Harvard ("SFFA") *increasingly treats attention to race itself as suspect. That creates a paradox for civil rights enforcement: institutions may need protected-class data to determine whether an ostensibly race-neutral system discriminates, even as collecting and analyzing those data may be characterized as impermissible, race-conscious decision making.*

This Article argues that civil rights law must distinguish race-conscious classification from race-conscious compliance. Classification uses race to determine how an individual will be treated. Compliance uses protected-class information to investigate, validate, audit, and monitor a facially neutral system without allowing race to determine an individual outcome. Existing doctrine already contains the foundations for this distinction, including

© 2026 Susan Tanner.

† Associate Professor at the University of Louisville Brandeis School of Law. I am grateful to JoAnne Sweeny, Joe Dunman, Julie Campbell, Dan Canon, and Cassie Chambers Armstrong for their thoughtful comments on an earlier draft, and to my research assistants, Katie Kimery and Taylor Griffith for their valuable research assistance. This project was supported by a research stipend from the University of Louisville Louis D. Brandeis School of Law.

disparate-impact doctrine, employment-testing law, Ricci v. DeStefano, *and Justice Kennedy's concurrence in* Parents Involved.

*The Article develops this distinction for algorithmic decision making and extends it to the harder problem of remediation after an audit reveals racial disparities. It proposes a protected, audit-safe harbor for institutions that separates demographic data from operational decision making, examines actual system performance, and uses race-conscious information to identify discrimination without assigning benefits or burdens by race. Without such protection, post-*SFFA *civil rights law risks creating an incentive for institutions not to measure discrimination at all. In an era when algorithms can reconstruct race without naming it, civil rights law must preserve the ability to see what formally colorblind systems hide.*

Table of Contents

I. Introduction 3
II. The Post-*SFFA* Problem: Formal Neutrality and Structural Blindness 5
III. How Algorithms Discriminate Without Race 12
IV. Merit, Fit, and the Cultural Construction of Predictive Targets 18
V. The Doctrinal Trap: Opacity, Causation, and Race-Conscious Auditing After *SFFA* 23
A. Identifying the Employment or Housing Practice 24
B. Statistical Disparity and Robust Causation 26
C. The Audit Paradox 28
D. Ricci Regulates Race-Based Action, not Investigation 30
E. Colorado and the Emerging Constitutional Challenge 31
F. The Doctrinal Line 33
VI. From Measurement to Remedy: Classification, Compliance, and Remediation 34
A. Classification Requires an Operative Decision Rule 35
B. Compliance Produces Knowledge Rather than Treatment 37
C. System-Level Remediation is not Necessarily Classification 39
D. Remediation Becomes Classification when Race Changes Individual Treatment 41
E. A Protected Audit Safe Harbor 43
F. Race-Conscious Purpose is not Race-Based Treatment 45

VII. Civil Rights Law Must Be Able to See what AI Hides ... 47

I. Introduction

The next civil rights fight over artificial intelligence will not begin with an algorithm that announces it is using race.[1] It will begin with a system that claims not to see race at all.

Employers already use automated tools to screen applicants. Landlords and lenders rely on algorithmic scoring systems. Schools use predictive analytics to identify students "at risk." Police departments deploy facial recognition and other surveillance technologies. Public agencies increasingly use automated systems to determine eligibility, prioritize cases, and allocate attention. These systems usually do not say that they are sorting people by race, sex, disability, national origin, or other protected traits. They speak in the language of prediction: risk, fit, likelihood, fraud, default, employability, safety, need, merit.

Those words are not technical descriptions; they are cultural choices. Civil rights law has long understood that facially neutral criteria can carry the preferences of the dominant group, whether the criterion is a paper credential, as in *Griggs v. Duke Power Co.*,[2] or a subjective judgment, as in *Watson v. Fort Worth Bank & Trust*.[3] Both cases rest on the same observation: a rule can be stated in entirely neutral terms and still operate as a sorting device, because the trait it measures was itself distributed unequally before the rule was written. Algorithmic systems do not dissolve that problem. They harden it. The criteria are now embedded in code, validated against historical data, and presented to applicants, regulators, and courts as objective measurement rather than as institutional preference.

---

1. This Article uses "AI" broadly to refer to automated systems, algorithmic processes, and machine-learning tools that materially assist or determine high-stakes decisions. The usage is deliberately functional rather than technical.

2. Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

3. Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 990–91 (1988) (plurality opinion).

Civil rights law is moving in the opposite direction. In *Students for Fair Admissions v. Harvard*,[4] the Supreme Court intensified its suspicion of racial classification. The Court treated the explicit use of race in university admissions as the central constitutional wrong, faulting the programs for pursuing goals too amorphous to measure, stereotyping applicants by racial category, operating as a negative for some applicants, and lacking any identifiable endpoint.[5] The opinion's most quoted formulation, "eliminating racial discrimination means eliminating all of it," compresses a formal account of equality into a constitutional command.[6] But algorithmic discrimination does not need explicit racial classification. A hiring model can learn from prior hiring patterns. A tenant-screening system can encode credit history, income volatility, criminal-system contact, eviction records, and neighborhood data. A school-risk model can reproduce unequal access to preparation, advising, attendance stability, and institutional support. A facial recognition system can impose unequal error burdens even when it is formally trained to identify faces, not race. The system can be "colorblind" in its inputs and still racially patterned in its outputs.

That gap between input and output creates a post-*SFFA* civil rights trap. The more the law treats race-consciousness itself as suspicious, the harder it becomes to detect discrimination in systems that hide racial hierarchy behind formally neutral data. Preventing algorithmic discrimination usually requires race-conscious auditing. An institution must ask whether its system produces different error rates, denial rates, rankings, recommendations, or burdens across racial groups; and it cannot ask that question without demographic data. Yet after *SFFA*, the same institutions may reasonably worry that collecting, analyzing, or acting on racial data will itself be read as the constitutional problem. The incentive that results runs in exactly the wrong direction. An institution that never measures cannot be shown to have known. By contrast, an institution that measures carefully

4. Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll., 600 U.S. 181, 214 (2023) [hereinafter *SFFA*].

5. *Id.* at 214–24.

6. *Id.* at 206.

generates the record on which a claim against it could be built. Legal doctrine that produces that incentive will get the behavior it rewards.

This Article argues that courts and agencies must distinguish between race-conscious classification and race-conscious compliance. Race-conscious classification uses race to allocate a benefit or burden to an individual. Race-conscious compliance uses race to test whether a facially neutral system is producing unlawful effects. *SFFA* speaks most directly to the first. Civil rights enforcement in the age of AI depends on preserving the second. The argument has three parts. First, algorithmic systems sort people through proxies and through culturally constructed targets like "merit" and "fit," producing racialized outcomes without racial classification. Second, post-*SFFA* doctrine threatens the auditing tools that would expose those effects. Third, the law should preserve a narrow form of race-consciousness for compliance even as it constrains race-conscious allocation. Without it, civil rights law will become most suspicious of the practices needed to reveal discrimination, while remaining too credulous toward systems that reproduce inequality without naming it. A legal regime that forbids institutions from seeing race cannot reliably govern technologies that make race harder to see.

II. The Post-*SFFA* Problem: Formal Neutrality and Structural Blindness

SFFA was about university admissions. It was not about employment testing, tenant screening, credit underwriting, insurance rating, or automated decision systems of any kind; and it did not address disparate impact, which was neither presented nor briefed.[7]

7. No opinion in *SFFA* addressed disparate-impact liability. At the time *SFFA* was decided, Title VI disparate-impact enforcement against recipients of federal financial assistance rested on implementing regulations, including 28 C.F.R. § 42.104(b)(2) (rescinded 2025) and 34 C.F.R. § 100.3(b)(2) (rescinded 2026), neither of which was before the Court. DOJ repealed 28 C.F.R. § 42.104(b)(2) in December 2025; and the Department of Education rescinded 34 C.F.R. § 100.3(b)(2) in July 2026. *See* Rescinding Portions of Department of Justice Title VI Regulations to Conform More Closely with the Statutory Text and to Implement Executive Order 14281, 90 Fed. Reg. 57141, 57141–42 (Dec. 10, 2025); Rescinding Portions of the Department of Education Title VI

footnote continued on next page

The holding is a good deal narrower than the opinion's rhetoric, and the distance between the two is where the trouble for AI governance begins. The Court asked whether Harvard and the University of North Carolina ("UNC") could use an applicant's race as a factor in evaluating that applicant for one of a fixed number of places and held that they could not.

The majority held that the admissions programs at Harvard and UNC were constitutionally impermissible because they relied on race in ways the Court found insufficiently disciplined. The programs' goals were too amorphous to permit judicial review because race operated as a negative for some applicants, racial categories risked stereotyping, and the universities had identified no meaningful endpoint.[8] The Court also warned that applicants' essays could not become a workaround for prohibited racial preferences: A university may consider what an applicant says about how race shaped his life but may not treat that discussion as a proxy for the preference the Court had just forbidden.[9]

Each of these defects presupposes a particular structure: an institution evaluating individual applicants for a limited number of places, with race operating as one input into that evaluation. Take them in order. To test whether a racial preference serves its goal, a court needs a measurable objective. That inquiry necessarily assumes a preference to test. Race operated as a negative because the places were finite, so that a benefit conferred on one applicant was necessarily a burden imposed on another. Stereotyping was a wrong done to the applicant being evaluated, who was treated as a representative of a category rather than as a person. The endpoint requirement asked when the preference would stop.[10] Now apply each of them to an employer that runs a regression to learn whether the screening tool it

Regulations to Align with the Statutory Text and Conform to Executive Order 14281, 91 Fed. Reg. 46733, 46733–42 (July 24, 2026).

8. *SFFA*, 600 U.S. at 214–24.

9. *Id.* at 230–31 (explaining that universities may consider an applicant's discussion of how race affected the applicant's life but may not use such discussion as a mechanism for racial preference).

10. *See id.* at 218–21, 226–30 (framing the negative-use, stereotyping, and endpoint problems in terms of their effect on individual applicants competing for admission).

licensed rejects Black applicants at a higher rate than white applicants. There is no applicant being evaluated by race, no scarce place being allocated, no individual asked to stand in for a category, and no preference to wind down. The regression tells the employer something about its own tool. It does not do anything to anyone.

For AI governance, the important development is not simply that affirmative action in admissions became constitutionally constrained. It is that the opinion reinforces a broader legal culture in which the use of racial categories is treated as the central civil rights danger. That culture is the latest expression of a long-running doctrinal contest between anti-classification and anti-subordination accounts of equality.[11] Anti-classification treats the government's use of race as the constitutional injury. Anti-subordination treats unjustified racial hierarchy as the injury, whether it is produced by classification, by neutral practices, or by institutional inertia. *SFFA* reads the Equal Protection Clause and Title VI through the anti-classification lens.[12] That reading makes intuitive sense when the challenged practice expressly allocates scarce opportunities by race, because the classification and the injury are the same thing. It makes considerably less sense when the question is whether a formally neutral system is covertly reproducing racial disadvantage, because there the classification is absent and the injury is not.

The logic does not follow, yet that extension is nonetheless being made; the effort has come principally from federal enforcement lawyers rather than from courts. Within two years of the decision, the Department of Education's Office for Civil Rights advised federally funded institutions that, although "*SFFA* addressed admissions decisions, the Supreme Court's holding applies more broadly," so that Title VI now reached race-conscious decisions in hiring, promotion, compensation, financial aid, scholarships, prizes, discipline, housing, graduation ceremonies, and "all other aspects of student, academic,

11. *See* Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIA. L. REV. 9, 11 (2003); Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over Brown*, 117 HARV. L. REV. 1470, 1539–40 (2004).

12. *SFFA* addressed both the Equal Protection Clause and Title VI of the Civil Rights Act of 1964, treating the two as imposing materially identical obligations on recipients of federal funding. *See SFFA*, 600 U.S. at 197–98.

and campus life."[13] Executive Order 14281 followed in April 2025, directing every federal agency to eliminate disparate-impact liability "to the maximum degree possible," withdrawing presidential approval of the Title VI disparate-impact regulations, and instructing agencies to begin repealing them. The Equal Employment Opportunity Commission ("EEOC") then closed its pending disparate-impact charges and adopted an enforcement plan that subordinates impact to intent.[14] None of this follows from *SFFA* as a matter of decided law. What *SFFA* supplied was the premise that any official attention to racial outcomes is presumptively something the Constitution forbids.

The Office of Legal Counsel completed the argument in June 2026. Responding to a request from the EEOC's Chair, the opinion concluded that the EEOC's Title VII disparate-impact guidelines are unconstitutional. It reasoned that "[b]y pressuring employers to take race-based actions in the name of proactively addressing potential statistical disparities, disparate-impact liability allows the government to engage in race discrimination indirectly."[15] It further concluded that the validation requirements of the Uniform Guidelines "radically deviate" from what Title VII's business-necessity defense demands, and that ordinary selection procedures should instead be presumed job-related.[16] The premises of that argument are worth stating plainly because they are the premises this Article contests. Measuring a statistical disparity is treated as the opening move of a race-based action. An institution's response to what the measurement shows is treated as continuous with the measurement itself. And the

13. Letter from Craig Trainor, Acting Assistant Sec'y for C.R., U.S. Dep't of Educ., to Colleagues 2–3 (Feb. 14, 2025), https://www.ed.gov/media/document/dear-colleague-letter-sffa-v-harvard-109506.pdf [https://perma.cc/KT4U-J4EY].

14. Exec. Order No. 14281, Restoring Equality of Opportunity and Meritocracy, 90 Fed. Reg. 17537 (Apr. 28, 2025); U.S. Equal Emp. Opportunity Comm'n, National Enforcement Plan for Fiscal Years 2025–2029 (June 4, 2026), https://www.eeoc.gov/sites/default/files/2026-06/NEP_-_signed.pdf [https://perma.cc/U82V-R2ED].

15. *Constitutionality of Disparate-Impact Liability Under Title VII*, 50 Op. O.L.C. __, 8 (June 9, 2026), https://www.justice.gov/olc/media/1444871/dl [https://perma.cc/4N5T-7TDT].

16. *Id.* at 19. The opinion also concluded that the Commission's voluntary affirmative action guidelines are unlawful. The Uniform Guidelines remain codified at 29 C.F.R. pt. 1607 (2026).

validation machinery that would allow an employer to demonstrate that its tool is lawful is recast as an instrument through which the government coerces racial decision making. Each step of that reasoning collapses a distinction this Article argues the law must keep.

The extension has not done well when tested. A federal district court vacated the Dear Colleague Letter in August 2025, holding among other things that it imposed new obligations without notice and comment and reached speech and curriculum that Title VI does not govern, and the Department dismissed its appeal the following January.[17] No court has held that an employer violates the Constitution by auditing a selection procedure for adverse impact. But litigation outcomes are not the mechanism that matters most here. Counsel advises clients on enforcement risk rather than on the eventual disposition of someone else's appeal, and an institution deciding whether to audit a vendor's hiring model will weigh an investigation, a document demand, and a press release more heavily than a favorable ruling three years later. That is the sense in which *SFFA* is coming to govern AI. Not as precedent, but as the source of a vocabulary in which the act of measuring racial effects can itself be described as the racial discrimination.

Justice Sotomayor and Justice Jackson each criticized the majority for treating racial inequality as though it could be made legally irrelevant by refusing to name it.[18] That disagreement is now central to AI governance, and the stakes are easy to state. If racial inequality is produced only by explicit racial classifications, then colorblind AI systems appear safe by construction. If racial inequality is also produced through proxies, histories, correlations, and institutional choices, then colorblind AI systems are precisely the ones that require scrutiny. Federal agencies have already recognized that automated systems can discriminate through biased training data, skewed design choices, opaque models, and variables correlated with protected traits.

---

17. Am. Fed'n of Tchrs. v. U.S. Dep't of Educ., 796 F. Supp. 3d 66, 123 (D. Md. 2025) (vacating the Dear Colleague Letter and related certification requirement; declining to vacate the FAQs because they were not final agency action), appeal dismissed, No. 25-2228 (4th Cir. Jan. 22, 2026).

18. *See SFFA*, 600 U.S. 181, 318-319 (2023) (Sotomayor, J., dissenting); *id.* at 388–90 (Jackson, J., dissenting). Justice Jackson did not participate in the Harvard case and wrote only in the UNC case.

In 2023, the Department of Justice ("DOJ"), the EEOC, the Consumer Financial Protection Bureau ("CFPB"), and the Federal Trade Commission issued a joint statement warning that automated systems can produce unlawful discrimination even when marketed as neutral or objective.[19] This statement identified three familiar pathways: data that reflect historical bias, models whose operation is difficult for anyone to understand, and deployment choices that create or magnify discriminatory effects.[20]

This agency position is still arguably applicable, even post-*SFFA*. Title VII reaches both intentional discrimination and facially neutral practices that produce unjustified disparate impact.[21] The Fair Housing Act reaches disparate impact under *Inclusive Communities*.[22] Title VI's implementing regulations historically prohibited disparate-impact discrimination by recipients of federal financial assistance. But the DOJ repealed its provision in December 2025, and the Department of Education rescinded its corresponding provision in July 2026.[23]

19. U.S. Dep't of Just. et al., Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems (Apr. 25, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf [https://perma.cc/YWH4-9AZP]; Lina M. Khan, *Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems*, FED. TRADE COMM'N (Apr. 25, 2023), https://www.ftc.gov/legal-library/browse/cases-proceedings/public-statements/joint-statement-enforcement-efforts-against-discrimination-bias-automated-systems [https://perma.cc/EAM8-N8QM].

20. *Id.*

21. 42 U.S.C. § 2000e-2(a), (k) (2018).

22. Tex. Dep't of Hous. & Cmty. Affs. v. Inclusive Cmtys. Project, Inc., 576 U.S. 519, 519 (2015).

23. At the time *SFFA* was decided, 28 C.F.R. § 42.104(b)(2) (rescinded 2025) and 34 C.F.R. § 100.3(b)(2) (rescinded 2026) prohibited certain practices with discriminatory effects. *See* Alexander v. Choate, 469 U.S. 287, 293 (1985) (describing Guardians as holding that unjustified disparate impacts may be redressed through Title VI implementing regulations); Alexander v. Sandoval, 532 U.S. 275, 293 (2001) (foreclosing a private right of action to enforce disparate-impact regulations). DOJ repealed 28 C.F.R. § 42.104(b)(2) (rescinded 2025) in December 2025, and the Department of Education rescinded 34 C.F.R. § 100.3(b)(2) (rescinded 2026) in July 2026. *See* Rescinding Portions of Department of Justice Title VI Regulations to Conform More Closely with the Statutory Text and to Implement Executive Order 14281, 90

footnote continued on next page

Regulation B likewise recognized an effects test until July 2026, when the CFPB amended the regulation to state that the Equal Credit Opportunity Act does not authorize disparate-impact liability.[24] Title VII and the Fair Housing Act continue to require effects-based analysis in disparate-impact cases. The former Title VI disparate-impact regulations are still relevant here because many of the entities they covered—universities, hospitals, school districts, and state and local agencies—are now buying automated systems in volume. A state child-welfare agency that licenses a risk-assessment tool, a public hospital system that adopts a triage model, and a school district that installs a predictive-analytics platform remain recipients of federal funding and thus remain subject to Title VI's prohibition on intentional discrimination. *SFFA*'s anti-classification logic, applied uncritically to these statutes, is in tension with the doctrinal architecture that makes them work. To know whether an automated system discriminates, an institution often must measure outcomes by protected class. It must ask whether Black applicants are rejected at higher rates, whether women are downgraded by a resume screener, whether disabled applicants are screened out by a video-interview tool, or whether facial recognition produces unequal error rates across racial groups.

These questions require race-conscious compliance. They do not require race-conscious classification. There is a substantial difference between using race to decide whether a student receives an admissions preference and using race to determine whether an algorithmic screening tool systematically disadvantages applicants of a particular race. The first changes how a particular person is treated. The second changes nothing about a particular person and instead evaluates whether an ostensibly race-neutral process is lawful. If courts collapse those categories, AI will acquire a new form of legal opacity. Institutions will be able to say truthfully that they do not use race,

Fed. Reg. 57,141 (Dec. 10, 2025); Rescinding Portions of the Department of Education Title VI Regulations to Align with the Statutory Text and Conform to Executive Order 14281, 91 Fed. Reg. 46,733 (July 24, 2026).

24. 15 U.S.C. § 1691 (2018); 12 C.F.R. § 1002.6(a) (2026) (stating that ECOA "does not provide that the 'effects test' applies"); Equal Credit Opportunity Act (Regulation B), 91 Fed. Reg. 21,620 (Apr. 22, 2026) (effective July 21, 2026).

while avoiding the very audits that would show whether their systems have reconstructed race through other variables.

III. How Algorithms Discriminate Without Race

Algorithmic discrimination often works through substitution. The system does not need race if it has access to variables that correlate with race or that encode racialized social conditions.[25] Removing the obvious proxies does not solve the problem because the information moves redundantly. A model denied zip code can often recover most of the same signal from commute time, school district, retail history, or the combination of several weak indicators that individually reveal nothing. The more variables a model has, the more paths there are between the data and the trait the institution believes it has excluded. A lender may use credit history, property valuation, employment stability, zip code, or consumer behavior. A landlord may use eviction records, criminal records, voucher status, or rental history. An employer may use resume features, employment gaps, educational pedigree, referral networks, or video-interview analytics. A school may use attendance, course-taking patterns, standardized test scores, disciplinary records, or signals of "engagement." Each of those variables can be defended as facially neutral. Many also have some genuine predictive value. That is part of the problem, rather than an answer to it.

Algorithmic systems rarely discriminate by using absurd or irrelevant variables. They discriminate by optimizing toward institutional goals using data drawn from a world structured by inequality, which means the more faithfully a model tracks the historical record, the more completely it absorbs the inequality in that record. *Griggs* anticipated the point fifty years ago. The high school diploma requirement and the general aptitude test at Duke Power were facially neutral and arguably useful, but they screened out Black applicants at substantially higher rates and bore no demonstrated relationship to job performance.[26] Title VII reaches such practices, the

25. *See* Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 Iowa L. Rev. 1257, 1273–76 (2020).

26. Griggs v. Duke Power Co., 401 U.S. 424, 425–28, 431–32 (1971).

Court explained, because the statute proscribes practices that are "fair in form, but discriminatory in operation."[27]

Algorithmic systems make that principle harder to administer. The challenged "practice" may be distributed across training data, feature selection, model architecture, thresholds, vendor defaults, and institutional deployment, with no single decision that a court can isolate and evaluate. The employer may not understand the tool. The vendor may claim proprietary protection. The applicant may never know why she was rejected, or that a model was involved at all. The Civil Rights Act of 1991 requires plaintiffs to identify particular practices producing disparate impact, unless the elements of the process are not capable of separation for analysis.[28] Algorithmic decision making presses hard against that requirement from both directions: The entire system may be the practice, but the system may also be too opaque for anyone outside it to dissect.

Housing illustrates the proxy problem, and it also illustrates how far the problem can shift from anything an institution would recognize as a decision. In the Department of Justice's Fair Housing Act case against Meta, the government alleged that Meta's housing-ad delivery system relied in part on protected characteristics in determining which users received which housing advertisements.[29] The mechanism is worth setting out. An advertiser selects an audience. But the platform decides which members of that audience actually see the advertisement, and it makes that decision by predicting engagement. If the historical data show that particular listings drew responses from particular users, the delivery system learns to route similar listings to

27. *Id.* at 431.

28. Civil Rights Act of 1991, Pub. L. No. 102-166, § 105, 105 Stat. 1071, 1074–75 (codified at 42 U.S.C. § 2000e-2(k)) (superseding, in relevant part, Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989)). The statute permits a plaintiff to challenge a decision making process as a whole if its elements are not capable of separation for analysis. 42 U.S.C. § 2000e-2(k)(1)(B)(i).

29. Press Release, U.S. Dep't of Just., Just. Dep't Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advert. (June 21, 2022), https://www.justice.gov/archives/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known [https://perma.cc/XGQ2-X92T].

similar users. An advertiser who selects an entirely neutral audience can therefore end up with a delivered audience skewed by race, without having asked for that result and without any way of observing it.

The remedy is as instructive as the allegation. The Department described the case as its first challenging algorithmic discrimination under the Fair Housing Act, and the settlement required Meta to stop using the audience tool at issue for housing advertisements and to build a replacement.[30] That replacement measures the difference between the audience eligible to see a housing advertisement and the audience that actually saw it, along with sex and estimated race or ethnicity, and brings that difference within a defined tolerance. An independent reviewer verified compliance, and the court supervised the arrangement until June 2026.[31] Nothing in that remedy required Meta to write race into a rule. Its objective was to keep race out of the delivered outcome. But the remedy could not have been designed, implemented, or verified without continuous measurement by race. The decree that resolved an algorithmic discrimination case was itself a race-conscious compliance regime.

Tenant screening presents a similar problem in a more decentralized setting. A landlord need not announce a racial rule if the screening score already incorporates credit history, rental history, eviction data, criminal records, or other variables produced by unequal housing and economic conditions. The resulting decision looks individualized, and from inside the transaction it is: this applicant, this score, this apartment. The effect across applicants can still be patterned by race; eviction records illustrate why. A filing appears in the record whether or not the tenant lost, whether or not the landlord was in the wrong, and whether or not the case was dismissed. Screening companies have also been found to over-include such records through automated matching procedures that attribute filings

30. *Id.*

31. Press Release, U.S. Dep't of Just., Just. Dep't and Meta Platforms Inc. Reach Key Agreement as They Implement Groundbreaking Resol. to Address Discriminatory Delivery of Hous. Advert. (Jan. 9, 2023), https://www.justice.gov/archives/opa/pr/justice-department-and-meta-platforms-inc-reach-key-agreement-they-implement-groundbreaking [https://perma.cc/2ZTY-7Z5K] (describing the replacement system, the variance metric, and the third-party review process).

to the wrong person, and the resulting entries are often neither corrected nor removed.[32] Filings are concentrated in the neighborhoods and buildings where filing is used as a routine collection tactic rather than as a last resort. A screening product that treats any filing as a negative signal therefore imports the filing practices of one housing market into every subsequent application the tenant makes. Because a small number of national vendors supply many landlords, a single scoring convention can propagate across an entire rental market, and the applicant who is turned away repeatedly has no way to learn that the same rule rejected her each time.[33]

Regulators have moved furthest in the domain of employment. New York City's Local Law 144 prohibits employers and employment agencies from using certain automated employment decision tools unless the tool has been subject to a bias audit within one year of use, a summary of the audit is publicly available, and required notices have been given.[34] The audit is not a general assessment of fairness; but its implementing rules require an independent auditor to calculate selection or scoring rates for each sex category, each race and ethnicity category, and the intersections of the two, and to express each as a ratio against the most-selected group.[35] The statute rests on a premise worth making explicit: An automated hiring tool cannot be evaluated

32. Consumer Fin. Prot. Bureau, Tenant Background Checks Market (Nov. 2022), https://www.consumerfinance.gov/data-research/research-reports/tenant-background-checks-market-report/ [https://perma.cc/HYQ6-G8T6] (finding that screening companies over-include eviction court records through automated matching procedures, and that the presence of an eviction record, without regard to its accuracy or outcome, is highly likely to result in denial).

33. Consumer Fin. Prot. Bureau, Consumer Snapshot: Tenant Background Checks (Nov. 2022), https://www.consumerfinance.gov/data-research/research-reports/consumer-snapshot-tenant-background-checks/ [https://perma.cc/994B-8C5G] (analyzing approximately 26,700 complaints submitted between January 2019 and September 2022); *see also* Consumer Fin. Prot. Bureau, Tenant Background Checks Market, *supra* note 32 (analyzing seventeen tenant screening companies).

34. N.Y.C. Admin. Code §§ 20-870 to 872 (enforcement effective July 5, 2023).

35. N.Y.C. Dep't of Consumer & Worker Prot., Rules Implementing Local Law 144, 6 R.C.N.Y. §§ 5-300 to 5-302 (defining the required bias audit and prescribing the calculation of selection rates, scoring rates, and impact ratios by sex category, race and ethnicity category, and their intersections).

without measuring its effects on protected classes, so an audit requirement is necessarily a requirement to gather demographic data.

The law's early record also shows what an audit mandate does not accomplish on its own. Because an employer decides for itself whether a given tool substantially assists or replaces discretionary decision making, the absence of a published audit does not establish that the law was violated. One study of 391 New York City employers found eighteen published audit reports, thirteen published transparency notices, and observed that nearly every published audit reported an impact ratio above four-fifths.[36] Those numbers are consistent with widespread noncompliance. They are also consistent with widespread good-faith conclusions that the law did not apply. The ambiguity is itself the finding. A disclosure regime that leaves the scope determination to the regulated party will produce measurement only where the party has already decided that measurement is required.

Federal guidance took the same view of the underlying obligation while it was in force. The EEOC's technical assistance on AI hiring applied the agency's longstanding Uniform Guidelines on Employee Selection Procedures, which set out the four-fifths rule and the obligation to validate selection procedures against actual job performance. Under that rule, a selection rate for any group that is less than eighty percent of the rate for the group with the highest selection rate generally indicates adverse impact.[37] The four-fifths ratio has never been a standard of liability. The Guidelines describe a shortfall as evidence that the enforcement agencies will generally regard as

36. Lucas Wright et al., *Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability*, 2024 PROCEEDINGS OF THE ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1701, 1701–13. The authors use "null compliance" to describe the condition in which the absence of a disclosure cannot be read as either compliance or its opposite.

37. U.S. Equal Emp. Opportunity Comm'n, Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964 (May 18, 2023) (withdrawn from the agency's website in 2025); 29 C.F.R. § 1607 (2026) (Uniform Guidelines on Employee Selection Procedures); *id.* § 1607.4(D) (four-fifths rule); *id.* § 1607.5 (validation standards).

adverse impact, which is a trigger for inquiry rather than a finding.[38] The distinction matters for algorithmic tools, because a vendor can tune a model until it clears four-fifths without ever addressing why the disparity was there. The guidance was withdrawn in 2025, but the withdrawal changes the enforcement signal not the statute. Further, Title VII's validation requirement does not depend on the agency's continued willingness to explain it.

Facial recognition adds a different kind of example, because the injury is an error rather than a denial. In its Rite Aid action, the Federal Trade Commission alleged that the company's facial recognition system generated false-positive matches that led employees to follow, detain, search, or publicly accuse customers—who had done nothing—and that those errors occurred more often in stores located in plurality-Black and plurality-Asian communities than in plurality-White ones.[39] The agency faulted Rite Aid for failing to consider and mitigate heightened risks tied to race or gender, and for failing to test, assess, measure, or document the technology's accuracy before deployment or to ask its vendor whether anyone else had done so.[40]

The remedy tracked the theory of the violation. The order bars Rite Aid from using facial recognition surveillance for five years and requires deletion of the biometric data and of any models derived from it, but the forward-looking provisions are the significant ones. If the company resumes such a program, it must test accuracy before deployment and monitor on a continuing basis whether the system's

38. 29 C.F.R. § 1607.4(D) (providing that a selection rate less than four-fifths of the rate for the group with the highest rate "will generally be regarded by the Federal enforcement agencies as evidence of adverse impact"). Courts have declined to treat the ratio as dispositive in either direction. *See* Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 995 n.3 (plurality opinion) (describing the four-fifths standard as a "rule of thumb").

39. Press Release, Fed. Trade Comm'n, Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology Without Reasonable Safeguards (Dec. 19, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without [https://perma.cc/VU28-2ZVZ].

40. *Id.*

errors fall disproportionately on people of a particular race or gender.[41] The remedy was not "do not classify by race," because the problem was never an explicit racial rule. The problem was the failure to test and govern a system whose errors were not evenly distributed. The civil rights harm could be seen only through measurement, which is to say both that an institution committed to not measuring would never have found it and that the agency's own remedy required the company to start.

IV. Merit, Fit, and the Cultural Construction of Predictive Targets

Every machine-learning model has a target variable: the thing the model is built to forecast. A hiring model predicts "good employees." A lending model predicts "creditworthy borrowers." A student-success model predicts "students likely to graduate." None of those targets exist as a natural fact waiting to be measured. Each must be operationalized, which means the institution chooses a measurable proxy and then treats that proxy as the truth the model is trying to recover. "Good employee" becomes "rated four or higher on the last performance review." "Creditworthy" becomes "did not default within two years." "Likely to graduate" becomes "graduated in six years from this institution under prior conditions."[42] The substitution disappears into the finished product, which reports a probability rather than the definitional choice that produced it. It is also the half of the problem that proxy analysis cannot reach. A model can be scrubbed of every variable correlated with race and remain aimed at a target that encodes it.[43]

Those are not technical choices; they are cultural and historical. Performance reviews reflect the preferences of the supervisors who

41. Stipulated Order for Permanent Injunction and Other Relief, FTC v. Rite Aid Corp., No. 2:23-cv-5023 (E.D. Pa. Mar. 8, 2024), https://www.ftc.gov/legal-library/browse/cases-proceedings/2023190-rite-aid-corporation-ftc-v [https://perma.cc/AM2R-T95M].

42. The locus classicus is Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671, 677–87 (2016) (describing how the choice and operationalization of target variables can produce discrimination).

43. *Id.* at 680–81 (distinguishing the choice of target variable from the choice of features, and observing that the former is the less visible of the two).

wrote them. Default histories reflect the lending practices of an industry with a documented record of redlining and discriminatory pricing.[44] Graduation rates reflect a campus environment that some students experienced as welcoming, and others did not. When a model is trained on labels of this kind, it does not learn what good employment, creditworthiness, or academic success actually are. It learns whom past institutions rewarded under past conditions. The output is presented as prediction. It is more accurately described as institutional memory, encoded.

The label is also a record of a particular moment. It fixes the judgments of the supervisors, underwriters, and admissions officers who were working when the data were generated, including judgments the institution has since revised or disavowed. An employer that overhauled its promotion criteria in 2020 and trained a model on a decade of prior outcomes has quietly reinstated the criteria it discarded. Nothing in the model announces that it is doing so, and the ordinary tests of model quality will not detect it, because those tests ask whether the model reproduces the historical record rather than whether the historical record deserves reproducing.

The choice of label is where a model's accuracy can be highest and its legal justification weakest. Consider a firm that defines a good employee as one who received a performance rating of four or higher. If supervisors at that firm rated Black employees lower than comparable white employees, a model trained on those ratings will learn the rating behavior along with everything else, and it will be measured as accurate precisely to the extent that it reproduces it. Improving the model makes the problem worse rather than better, because a more faithful model tracks the biased label more closely. Validation against the label will confirm that the model works. Validation against the job will not have been attempted. The two are

44. *See generally* Richard Rothstein, The Color of Law: A Forgotten History of How Our Government Segregated America (2017) (documenting federal and private redlining practices and their effects on housing wealth).

easy to confuse because both are described as showing that the model predicts performance.[45]

Griggs rejected the diploma requirement and the general aptitude test at Duke Power because the employer had never validated either against the work itself.[46] The company adopted the requirements without any study of whether they predicted performance in the jobs at issue, and the record showed that employees hired before the requirements existed were performing those jobs satisfactorily.[47] The Court's objection was not that the credential was irrational; it was that no one had checked, and that the burden of not checking fell on Black applicants.[48]

The principle went from credentials to judgments. In *Watson*, the Court held that subjective evaluation criteria such as supervisor ratings of "leadership," "interpersonal skills," and "fit" can be challenged under disparate-impact theory rather than being insulated as matters of managerial discretion.[49] Justice O'Connor's plurality reasoned that "an employer's undisciplined system of subjective decision making" can produce precisely the same patterned exclusion as a written rule, and that a contrary holding would invite employers to escape the statute by declining to write their criteria down.[50] The Uniform Guidelines embed the same insight in administrative form: A selection procedure that produces adverse impact must be validated, and the validation must connect the criterion to actual job requirements rather than to the institution's preferences.[51]

Predictive models are a form of subjective evaluation in technical clothing. The choice of target variable is a choice about what counts. A model that predicts "employee retention" privileges fit with the

45. *See* Barocas & Selbst, *supra* note 42, at 678–80 (explaining that a model trained on a biased label will be scored as accurate to the extent it reproduces the bias).

46. Griggs v. Duke Power Co., 401 U.S. 424, 433 (1971).

47. *Id.* at 427, 431–32.

48. *Id.* at 431 ("The touchstone is business necessity."); *see also id.* at 432 (placing on the employer the burden of showing that a requirement bearing a disparate impact is related to job performance).

49. Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 989–91 (plurality opinion).

50. *Id.* at 990–91.

51. 29 C.F.R. § 1607.5(C) (2026) (validation standards).

existing workforce, since the employees who stayed are the ones who defined the label. A model that predicts "promotion within five years" privileges patterns that may reflect mentorship networks, sponsorship, and access rather than capacity. A model that predicts "engagement" privileges the behaviors that already look like engagement to the institution. None of these targets is wrong on its face. None is neutral, either. Each is a cultural artifact dressed as a number.

Watson's concern was that discretion left unwritten escapes review.[52] A target variable is discretion written down and then compiled, which makes it more reviewable in principle and less reviewable in practice. The judgment is now recorded with unusual precision, in a specification any auditor could read. But the applicant cannot read it; the employer that licensed the model may never have seen it; and the vendor that wrote it treats it as proprietary. The doctrine assumes that reducing a judgment to a rule makes the judgment easier to test. Automation reduces the judgment to a rule and simultaneously removes it from view.

Lani Guinier described American higher education's reliance on standardized assessment as a "testocracy" that confuses the measurement of preparation with the measurement of capacity.[53] Devon Carbado and Mitu Gulati documented how informal criteria like "professionalism" and "fit" function as workplace gatekeeping that imposes additional costs on workers of color, who must exert effort adopting an identity that some of their colleagues simply have by default.[54] Solon Barocas and Andrew Selbst showed how each step of the data-mining process, from problem definition to target variable selection, can encode the discrimination that civil rights law was designed to reach.[55] Pauline Kim has argued that data-driven employment decisions fit awkwardly into Title VII precisely because they distribute decision making across many ostensibly neutral

52. *See Watson*, 487 U.S. at 990–91 (plurality opinion).

53. Lani Guinier, The Tyranny of the Meritocracy: Democratizing Higher Education in America 26–43 (2015).

54. Devon W. Carbado & Mitu Gulati, *Working Identity*, 85 Cornell L. Rev. 1259, 1273–82 (2000).

55. Barocas & Selbst, *supra* note 42, at 677–93.

choices, none of which looks like the discrete practice the statute contemplates.[56] The four accounts converge on a single point: The criterion that appears neutral is doing the work that an explicit rule would have done, and it is doing that work in a form the law is not accustomed to examining.

Under Title VII, an employer using an algorithmic hiring tool with adverse impact must show that the tool is "job-related for the position in question and consistent with business necessity."[57] Predictiveness against a target variable is not the same thing as job-relatedness. If the target is itself a measure of past institutional preferences, the model can be highly predictive and still legally inadequate, because it will have predicted the preference rather than the work. A defense built on accuracy figures, therefore, proves less than it appears to prove. It establishes that the model recovers the label, which was never the contested question. Validation must connect the model's output to a defensible account of what the job requires, not to a self-fulfilling record of who succeeded under prior conditions.[58]

The merit problem also reframes the *SFFA* debate. The opinion treats race-conscious admissions as the deviation from a neutral baseline of academic merit. But the baseline of "merit" in selective admissions is itself a constructed target: a composite of standardized tests, grades earned in unequally resourced schools, recommendations written within unequal mentorship networks, and extracurricular profiles shaped by family resources.[59] None of those inputs is racially neutral in its production, even though none is racially classified in its use. An algorithmic admissions tool trained on past admissions decisions would inherit the same structure, and would inherit it

56. Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 883–910 (2017).

57. 42 U.S.C. § 2000e-2(k)(1)(A)(i).

58. *See* 29 C.F.R. § 1607.5(B) (2025) (requiring that validity studies show that the selection procedure measures important work behaviors or outcomes); *cf.* Albemarle Paper Co. v. Moody, 422 U.S. 405, 431–32 (1975) (rejecting a validation study that did not connect the tested attributes to the requirements of the jobs at issue).

59. *See* GUINIER, *supra* note 53, at 24–47; *see also SFFA*, 600 U.S. at 388–96 (Jackson, J., dissenting) (describing the historical and structural conditions that shape the inputs to academic merit).

invisibly, because the tool would report a score rather than a set of judgments. The question is not whether the system uses race. The question is whose conception of merit it has automated.

If the law treats only explicit racial classifications as suspect, it will treat the choice of target variable as a neutral institutional prerogative. But that choice is where most algorithmic discrimination is produced. A model can be entirely race-blind in its inputs and still operationalize a culturally bounded conception of merit. Civil rights law has the doctrinal resources to address this, through *Griggs*, *Watson*, and the validation requirements of the Uniform Guidelines.[60] What it lacks, in the post-*SFFA* environment, is permission to use those resources fully.

V. The Doctrinal Trap: Opacity, Causation, and Race-Conscious Auditing After *SFFA*

Disparate-impact doctrine reaches facially neutral practices that produce unjustified group disparities, and it does so without requiring proof of anyone's state of mind. That makes it the natural home for algorithmic discrimination claims, because it is suited to discrimination embedded in an institutional system rather than announced in an individual decision. An algorithm need not harbor intent. It can reproduce historical exclusion, reward traits associated with the institution's existing workforce, or distribute errors unequally across protected groups; and it can do all of that while every person operating it acts in good faith. Disparate-impact doctrine asks what the practice does and whether the practice can be justified.

The doctrine assumes, however, that a plaintiff can identify a practice, measure its effects, connect the practice to the disparity, and evaluate the defendant's justification. Algorithmic systems make each of those steps unusually difficult. The relevant "practice" may be dispersed across data collection, label construction, feature selection, model training, threshold setting, vendor defaults, and the way institutional employees actually use the resulting score. The applicant ordinarily sees none of those choices. She sees only the consequence: no interview, no apartment, no loan, no admission, with no explanation. That opacity creates two related problems. The first is an

60. *See supra* notes 46–51 and accompanying text.

enforcement problem. A plaintiff may not possess enough information to plead or prove that a particular component caused the disparity, and the information she would need is held by the defendant. The second is a compliance problem, and it operates independently of litigation. An institution cannot determine whether its own system is discriminatory unless it has access to the same information, and much of that information sits with a vendor. A doctrine that demands precise causal proof while permitting developers and deployers to conceal the information needed to produce that proof risks making algorithmic discrimination legally unrecognizable.

A. Identifying the Employment or Housing Practice

Title VII requires a plaintiff to identify a "particular employment practice" that causes a disparate impact. When the elements of the decision-making process cannot be separated for analysis, however, the statute permits the process to be analyzed as a single employment practice.[61] Congress added that provision in 1991 in response to a decision that had placed the burden of disaggregation on plaintiffs, and its premise is that a plaintiff should not lose because the defendant's process is complicated.[62] The provision is especially important for algorithmic systems. The operative practice may not be one variable or one line of code. It may be the interaction among the model, the data, the decision threshold, and the institutional setting in which the output is used.

Consider an employer that purchases a candidate-screening system from a vendor. The vendor trains the model using data drawn from prior employees. The employer selects "retention after two years" as the desired outcome. The model assigns applicants a score, but the employer chooses the cutoff below which applicants will not receive human review. Recruiters are permitted to override the score, although no one records how often they do so. If the resulting system excludes Black applicants at a substantially higher rate, what is the challenged practice? It could be the training data, the retention label, the features the model treats as predictive, the threshold selected by

61. 42 U.S.C. § 2000e-2(k)(1)(A)–(B) (2024); *see* Lewis v. City of Chicago, 560 U.S. 205, 212–14 (2010).

62. *See supra* note 28.

the employer, the absence of review, or the complete decision-making process. Each candidate answer is plausible, and no one outside the firm can choose among them.

The difficulty is not that the question is unanswerable. It is that answering it requires access to the vendor's training data, the employer's configuration choices, and the recruiters' override logs, none of which the applicant has and at least one of which may not exist. The statutory requirement to identify a practice should not become a demand that a plaintiff reverse engineer a system before discovery. In a conventional testing case, an applicant knows that a written examination screened her out, and she can point to the examination. In an algorithmic case, the applicant may not know that a model was used at all, much less which component produced the result. The entity controlling the system is also the entity controlling the evidence needed to isolate its effects, so the informational asymmetry runs in only one direction. Courts should therefore permit plaintiffs to challenge the integrated process when the defendant has not made its components meaningfully available for separate analysis. Read otherwise, the statutory whole-process provision would protect only those systems simple enough for outsiders to understand, which is the opposite of the category it was written to reach.

Nor should a favorable aggregate outcome answer the question. *Connecticut v. Teal* rejected the argument that an employer may defend a discriminatory component by pointing to a nondiscriminatory bottom line, holding that Title VII protects individuals against barriers erected in the course of a selection process and not merely against unfavorable overall statistics.[63] The principle is directly applicable to a staged automated process. A screening model that eliminates Black applicants at a substantially higher rate is not cured because recruiters later restore the overall proportion through manual review, any more than the written examination in *Teal* was cured by the promotions that followed it.[64]

A vendor may argue that it supplied a general-purpose tool and did not control the employer's target, threshold, or deployment. The

63. Connecticut v. Teal, 457 U.S. 440, 450–56 (1982).

64. *See id.* at 442, 453-54 (rejecting the argument that a nondiscriminatory "bottom line" precludes a prima facie case or supplies a defense).

employer may argue that it reasonably relied on the vendor's representations and lacks access to the model's internal operation. Both claims may contain some truth, and together they hold no one accountable for the decision making process. But contractual fragmentation should not produce a civil rights vacuum. The employer remains responsible for an employment practice it uses, even when a third party supplied the technology, and federal guidance has long taken that position with respect to vendor-supplied tests.[65] The vendor's role may present separate questions of liability, contribution, discovery, or regulatory responsibility. But outsourcing a selection procedure does not make the procedure cease to be an employment practice.

B. Statistical Disparity and Robust Causation

In *Inclusive Communities*, the Supreme Court held that a statistical disparity, standing alone, does not establish disparate impact under the Fair Housing Act. A plaintiff must identify a policy that caused the disparity. The Court described this as a "robust causality" requirement, designed to prevent liability from attaching to racial imbalance that the defendant did not create, and it warned that remedies should focus on eliminating the offending practice and should generally proceed through race-neutral means.[66] Those limits serve a real purpose. They protect legitimate policies from being displaced merely because social inequality appears somewhere in the background, and they keep disparate-impact liability from becoming a roving license to equalize outcomes.

A model may rely on hundreds of variables, on interactions among variables, and on correlations that cannot be translated into a single narrative explanation. The system may also change over time as its data, its users, or the surrounding institutional practices change, so that the model that produced last year's disparity is not the model

65. *See Employment Tests and Selection Procedures*, U.S. EQUAL EMP. OPPORTUNITY COMM'N (Dec. 1, 2007), https://www.eeoc.gov/policy/docs/factemployment_procedures.html [https://perma.cc/Y5GX-A4ND] (explaining that employers remain responsible for ensuring that vendor-supplied tests comply with federal law); 29 C.F.R. § 1607.7(A) (2025).

66. Tex. Dep't of Hous. & Cmty. Affs. v. Inclusive Cmtys. Project, Inc., 576 U.S. 519, 542–43 (2015).

available for inspection today. A disparity may arise from no single feature. It may arise from the model's aggregation of many individually weak signals that collectively reconstruct race, sex, disability, or class position. Robust causation, applied to a system of that kind without adjustment, would function as an immunity rule rather than a limiting principle.

Robust causation should not be confused with monocausal explanation. Civil rights plaintiffs need to connect the challenged system to the observed disparity, but they should not be required to identify one discriminatory variable when the system produces its output through interaction. The relevant causal question is whether use of the system materially changes protected-group outcomes compared with an appropriate baseline, not whether one feature can be designated the discriminatory feature.[67] An institution should not be able to defeat causation by dividing one integrated decision among enough technical components.

Baselines are easy to get wrong in ways that hide the model's contribution. Comparing the model's output with the final institutional outcome may obscure that contribution if human reviewers routinely override some recommendations and not others. Comparing accepted applicants with the entire labor market may obscure disparities introduced earlier by targeted advertising or recruiting, before the model ever ran. A serious audit should therefore examine each stage of the process and the total process, an approach the Court has already required in the analogous setting of multistage employment testing.[68] It should ask who enters the pool, who advances, who receives human review, whose scores are overridden, and who ultimately receives the opportunity.

Auditing makes that inquiry more disciplined rather than more speculative. An audit can compare outcomes with and without the system, examine selection rates at different stages, test alternative thresholds, and calculate error rates across protected groups. It can help determine whether a disparity arises from the applicant pool, the model, the deployment decision, or some combination of them.

67. *See Cf. Inclusive Communities*, 576 U.S. at 542–43.

68. *See Teal*, 457 U.S. at 450–56; *see also* 29 C.F.R. § 1607.4(C) (2025) (addressing analysis of the components of a selection process).

Properly conducted audits therefore support the caution *Inclusive Communities* demanded rather than undermining it. They are the mechanism by which a disparity caused by the challenged system can be distinguished from a disparity attributable to conditions the defendant did not create.[69]

The alternative is not a genuinely colorblind causal inquiry. It is a causal inquiry deprived of the information necessary to reach a conclusion. Preventing the collection or analysis of protected-class data would not make causal findings more reliable. It would make them less testable, which is a different thing and a worse one.

C. The Audit Paradox

Civil rights law has never treated protected-class data as inherently unlawful. Title VII enforcement depends on employers recording enough information to determine whether their selection practices exclude racial, ethnic, or sex groups at different rates. The Uniform Guidelines instruct covered users to maintain records showing the impact of their selection processes on identifiable groups, and the Guidelines' familiar four-fifths rule cannot be applied at all without group data.[70]

The rule compares the selection rate for each group with the selection rate for the group selected at the highest rate, and treats a ratio below eighty percent as evidence that the enforcement agencies will generally regard as adverse impact.[71] Every term in that comparison is a group rate. An employer that does not know which applicants belong to which group cannot compute a single one of them and therefore cannot determine whether the rule it has been following for four decades has been violated. An employer in that position is not

69. *Cf. Inclusive Communities*, 576 U.S. at 542 (describing the "robust causality" requirement as a safeguard against liability premised on racial imbalance alone).

70. 29 C.F.R. §§ 1607.4, 1607.15 (2025); Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, Questions 11–13, 85–88, 44 Fed. Reg. 11,996, 11998, 12008 (Mar. 2, 1979).

71. 29 C.F.R. § 1607.4(D) (2025). The ratio is framed as evidence the enforcement agencies will generally regard as adverse impact, not as a definition of liability.

neutral with respect to adverse impact. It is unable to form a view about it.

The separation of demographic data from operational decisions is what makes this workable. An employer can know an applicant's race for compliance purposes without providing that information to the model or to the decisionmaker. Protected-class information may be stored in a separate system, subject to limited access, and used only after decisions have been generated. The employer can then compare selection rates and error rates without race having entered any applicant's score. That is not a racial classification governing treatment. It is a method of testing whether the actual selection procedure complies with law.

The same structure applies before deployment. Developers may need protected-class data during validation to compare candidate models, to identify unequal error rates, or to determine whether an apparently accurate feature creates substantial group disparities. What matters is not whether protected information was ever considered. What matters is whether it becomes an operative rule for assigning treatment to an individual.

A model audit comparing false-positive rates across racial groups does not impose a false positive on anyone because of race. A validation study identifying whether a hiring tool excludes Black applicants at a higher rate does not hire or reject any applicant because of race. A lender's comparison of denial rates does not alter any applicant's creditworthiness. These activities produce knowledge about the system. They do not allocate benefits or burdens. The decisions they evaluate have, in every case, already been made.

That does not make demographic data harmless. Protected-class information is sensitive. It can be exposed, repurposed, or combined with other data in ways that affect the very people whose information made the audit possible. An audit regime should therefore require data minimization, access controls, retention limits, documentation, and separation from operational decision making. Those safeguards respond to the real risk, which is that information gathered for compliance will migrate into individual treatment. The possibility of misuse justifies governing the information. It does not justify institutional ignorance.

D. Ricci *Regulates Race-Based Action, not Investigation*

Ricci v. DeStefano appears at first to complicate this analysis. New Haven, Connecticut, declined to certify firefighter-promotion examination results after the results showed substantial racial disparities. The Supreme Court held that the city's action violated Title VII because the city lacked a strong basis in evidence that certifying the results would have subjected it to disparate-impact liability. The decision reconciled Title VII's disparate-treatment and disparate-impact provisions by limiting the circumstances in which an employer may take race-based action against identifiable individuals in order to avoid anticipated disparate-impact liability.[72]

The important phrase is "race-based action." New Haven did more than measure the test's effects. It discarded completed results and denied particular candidates the opportunity to compete for promotions because of the racial distribution of those results. The decision changed the legal consequences of the test for identifiable individuals, each of whom had already taken it and some of whom had prepared for months. Section 703(l) of Title VII similarly prohibits an employer from adjusting scores, using different cutoffs, or otherwise altering employment-test results on the basis of race.[73] Both provisions concern the allocation of employment opportunities after race has entered the decision rule.[74]

Ricci did not hold that the city acted unlawfully when it collected the candidates' racial data, calculated disparate impact, asked experts to examine the test, considered validation evidence, or investigated alternative testing methods. The Court's own recitation of the facts describes those steps at length and without disapproval.[75] All of them occurred before the city declined to certify the results, and the Court did not characterize any of them as disparate treatment. Its objection was to the city's eventual decision to discard the results without sufficient evidence that doing so was necessary to avoid liability.

---

72. Ricci v. DeStefano, 557 U.S. 557, 579–85 (2009).

73. 42 U.S.C. § 2000e-2(l) (2024).

74. *See id.*; *Ricci*, 557 U.S. at 584–85.

75. *See Ricci*, 557 U.S. at 562–74 (recounting the city's collection of the results, its consultation with the examination developer and other experts, and its consideration of alternative selection methods).

The case therefore supports a boundary between diagnosis and remedy. An employer does not need a strong basis in evidence before it investigates whether a selection procedure has adverse impact. Requiring such evidence would be circular: The employer would need proof of likely liability before undertaking the analysis by which that proof could be discovered. The strong-basis requirement applies when the employer moves from examining racial effects to changing an already completed employment outcome on the basis of race.[76]

Suppose an audit reveals that two models predict job performance with approximately equal accuracy, but one produces substantially less racial disparity. Selecting the less discriminatory model does not require the employer to give any applicant a preference because of race. Every future applicant is evaluated under the same model, and no applicant's score depends on her race. The employer's reason for choosing the model may be race conscious, but the operational decision rule is not. Likewise, an employer may remove a proxy variable, revise a poorly supported target, lower the weight assigned to a credential, or add human review for low-confidence predictions without classifying applicants by race. What distinguishes these measures from New Haven's is that they operate prospectively on a rule rather than retroactively on a result.

That distinction does not resolve every case. A purportedly neutral redesign could be a pretext for constructing outcomes according to racial targets. A system could also impose formally uniform rules that have been engineered to guarantee a predetermined racial distribution, which would be a classification wearing a neutral rule as a costume. Courts should examine what the institution changed and how the revised system treats individuals. But awareness of group effects should not itself transform a generally applicable design change into racial discrimination.

E. Colorado and the Emerging Constitutional Challenge

In 2024, Colorado enacted Senate Bill 24-205, which required developers and deployers of certain high-risk AI systems to exercise reasonable care to protect consumers from algorithmic discrimination. The law covered consequential decisions in education,

76. *Id.* at 585.

employment, financial services, essential government services, housing, insurance, and other high-stakes domains.[77] Its central obligation was a duty of care rather than a prohibition, which meant that a covered firm satisfied the statute by managing a risk rather than by achieving any particular distribution of outcomes. But managing a risk of discriminatory effect presupposes an ability to observe that effect and observing it requires demographic data. The statute thus imposed, as a practical matter, an obligation to measure.

In April 2026, xAI challenged the statute in federal court. The United States moved to intervene and argued, among other things, that the statute violated equal protection by requiring companies to prevent unintended effects associated with protected characteristics while permitting some conduct intended to promote diversity or redress historical discrimination.[78] The complaint characterized the duty to prevent disparities as a demand that AI developers alter otherwise neutral outputs according to race or sex.[79] The step from a duty to measure to a demand to alter is the whole of the argument, and it is made rather than defended.

Colorado then repealed and reenacted the statute through Senate Bill 26-189, signed on May 14, 2026. The replacement regulates automated decision-making technology used to make or materially influence consequential decisions. It places greater emphasis on developer documentation, consumer notice, explanations following adverse outcomes, correction of inaccurate personal data, meaningful human review, and record retention. It also addresses how responsibility is allocated between developers and deployers in civil actions brought under existing antidiscrimination law. Most of the new provisions take effect January 1, 2027; specified disclosure,

77. An Act Concerning Consumer Protections in Interactions with Artificial Intelligence Systems, 2024 Colo. S.B. 205.

78. U.S.'s Complaint in Intervention, X.AI LLC v. Weiser, No. 1:26-cv-01515 (D. Colo. Apr. 24, 2026); Press Release, U.S. Dep't of Just., Justice Department Intervenes in xAI Lawsuit Challenging Colorado's 'Algorithmic Discrimination' Law (Apr. 24, 2026), https://www.justice.gov/opa/pr/justice-department-intervenes-xai-lawsuit-challenging-colorados-algorithmic-discrimination [https://perma.cc/5PHX-KZJH].

79. Complaint in Intervention, X.AI LLC v. Weiser, No. 1:26-cv-01515 (D. Colo. filed Apr. 24, 2026).

consumer-rights, and enforcement provisions took effect upon enactment.[80]

The sequence should be described carefully. The replacement of Senate Bill 24-205 does not itself establish that the constitutional challenge caused Colorado to change its law, and no court has held that a protected-class audit violates equal protection. The episode nevertheless exposes the conflict this Article identifies. A regulatory obligation directed at detecting and reducing protected-class disparities was characterized by challengers as an unconstitutional demand that AI systems be altered according to race.

The replacement statute also does not eliminate the underlying problem. Notice and human review may improve procedural accountability, and a consumer who learns that an automated system contributed to an adverse decision is better positioned than one who does not. But neither requirement reveals whether the system produces systematically different selection rates or error burdens across racial groups, because neither aggregates. A thousand individually accurate notices disclose nothing about the pattern they collectively describe. Allocating responsibility between developer and deployer cannot operate effectively either if neither party is expected to measure the system's protected-class effects, since the allocation presupposes an injury that no one has been asked to detect. The need for race-conscious diagnosis remains even when the regulatory vocabulary shifts away from algorithmic discrimination.

F. The Doctrinal Line

The doctrinal line should turn on what race does within the challenged practice. When race determines an individual's score, cutoff, eligibility, or treatment, the ordinary rules governing racial classification, disparate treatment, and race-based remedies apply. When race is held apart from the operational system and used to measure whether a facially neutral process produces unlawful disparities, the activity is compliance rather than classification.

That line can be applied through questions courts already know how to ask. Did protected-class data enter the individual decision? Did

80. An Act Concerning the Use of Automated Decision-Making Technology in Consequential Decisions, 2026 Colo. S.B. 189.

decisionmakers have access to it? Did the audit change treatment for particular people? Does any redesign apply under the same terms to all individuals? Did the institution use racial targets as a substitute for evaluating the legality of the underlying practice? Each question addresses how race operates within the system rather than treating any awareness of race as dispositive, and each is answerable from documents an institution running a lawful audit will already have generated.

Without that distinction, civil rights law creates a one-way ratchet. Institutions may adopt opaque systems that reproduce racial hierarchy through proxies and historical data, but efforts to uncover those effects become legally suspect because they require naming race. The result is not neutrality. It is an evidentiary privilege for discrimination that can be expressed without racial terminology.

VI. From Measurement to Remedy: Classification, Compliance, and Remediation

The distinction between race-conscious classification and race-conscious compliance is clearest at the endpoints. A university classifies by race when an applicant's race changes how the university evaluates the application. An employer engages in race-conscious compliance when it stores demographic information separately from applications and later uses that information to determine whether a selection procedure has adverse impact. The first practice alters individual treatment. The second produces information about whether a facially neutral process operates lawfully.

The harder cases arise between those endpoints. Once an audit identifies a disparity, the institution must decide whether and how to respond; doing nothing is itself a decision with legal consequences. It may revise the model's target, remove a proxy variable, change a threshold, select an alternative model, add human review, or abandon the automated system altogether. Those decisions may be motivated by knowledge of racial effects even when the revised procedure applies identically to every person. At some point, however, a purported system redesign may become an effort to allocate outcomes according to race.

The law therefore needs three concepts rather than two. Race-conscious classification uses race as an operative rule in deciding how an individual will be treated. Race-conscious compliance uses protected-class information to investigate, measure, validate, or monitor a system without allowing that information to determine an individual outcome. Race-conscious remediation occupies the space between them. It uses information generated through compliance to alter the system. Some forms of remediation remain system-level and facially neutral. Others introduce race into individual decisions and become classifications.

Remediation is not a third constitutional category with its own level of scrutiny. It is the point at which courts must determine whether an institution has moved from observing racial effects to assigning treatment by race. The question is not merely whether the institution was aware of race or hoped to reduce a racial disparity. The question is what role race plays in the decision rule that governs individuals. Naming remediation separately matters because it is where most real disputes will arise. Few institutions will be caught writing race into a threshold, and few will be sued for running an audit. The contested cases will involve an institution that measured, found something, changed the system, and must now explain what it changed and why.

A. Classification Requires an Operative Decision Rule

Equal protection doctrine is most demanding when the government assigns a benefit or burden according to an individual's racial category. *Adarand*, *Croson*, *Parents Involved*, and *SFFA* each involved a program in which race affected the allocation of a contract, a school assignment, or an admissions opportunity.[81] The constitutional concern in those cases was not that officials possessed demographic information or understood that their policies had racial consequences. Every one of those programs was designed by officials who had studied racial data closely, and the Court did not suggest that

81. *SFFA*, 600 U.S. 181, 206–24 (2023); Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720–25 (2007); Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 205–10 (1995); City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493–98 (1989).

the studying was the problem. It was that an individual's racial designation had become part of the rule governing that individual's treatment.

Parents Involved makes the point directly. Justice Kennedy supplied the fifth vote to invalidate school-assignment plans that used an individual student's race as a tiebreaker.[82] But his concurrence also rejected the proposition that the Constitution requires government to disregard racial conditions. School systems, he explained, could track enrollment and performance by race, recruit students and faculty in targeted ways, allocate resources among schools, and adopt general policies designed with knowledge of racial demographics.[83] What those measures had in common was that none of them assigned a particular child to a particular school because of that child's race. They were race conscious without being classifications, and Justice Kennedy treated the difference as constitutionally decisive rather than as a matter of degree.

The admissions programs in *SFFA* did not merely study whether their existing processes produced racial disparities. They used race in evaluating applicants for a limited number of places. The Court described the constitutional problem repeatedly in terms of classifying applicants, providing preferences, imposing disadvantages, and using race as a factor in individual admissions decisions. *SFFA* strengthens the prohibition against racial allocation. It does not establish that every institutional action taken with awareness of racial consequences is itself a racial classification, and it did not purport to overrule the line Justice Kennedy drew.[84]

For algorithmic systems, race becomes an operative decision rule when the system uses an individual's racial category to determine a score, threshold, ranking, recommendation, or outcome. The test is not whether race appears somewhere in the institution's files. It is whether changing a single applicant's recorded race, holding everything else constant, would change what the system returns for that applicant. A

82. *Seattle Sch. Dist. No. 1*, 551 U.S. at 782–783.

83. *Id.* at 788–89 (Kennedy, J., concurring in part and concurring in the judgment).

84. *See SFFA*, 600 U.S. at 218–30; *cf. Parents Involved*, 551 U.S. at 788–89 (Kennedy, J., concurring in part and concurring in the judgment).

lender that applies a different approval threshold to Black and White applicants classifies by race. An employer that adds points to the scores of applicants from one racial group classifies by race. A government agency that routes otherwise similar people into different eligibility procedures according to race classifies by race. The fact that a computer performs the operation rather than a human does not alter the legal character of the decision.

The same concern arises when an institution does indirectly what it could not do directly. A model designed to infer an individual's race and then use the inferred category as an allocation rule should not escape scrutiny merely because race does not appear as a labeled field in the data.[85] Nor should an institution avoid review by assembling a collection of proxies chosen for the purpose of approximating racial identity. The relevant question is functional: Does the system sort individuals for different treatment because they belong, or are predicted to belong, to a racial group?

B. Compliance Produces Knowledge Rather than Treatment

Compliance treats protected-class information as an evaluative variable rather than an operational one. Race is used to ask whether the system produces materially different outcomes, errors, or burdens across groups. It is not used to decide which outcome a particular person receives. The difference is not one of degree or intensity; it is a difference in what the information is permitted to touch.

The Uniform Guidelines embody that distinction. They expect employers to calculate selection rates across racial, ethnic, and sex groups to determine whether a process produces adverse impact and to maintain demographic records sufficient to conduct that analysis.[86] Federal guidance also contemplates telling applicants that demographic information is requested for recordkeeping, and not for employment decisions, which is the same separation stated as a disclosure.[87] The federal enforcement framework could not function if

85. *See* Prince & Schwarcz, *supra* note 25, at 1268–80 (describing how models can reconstruct protected characteristics from unprotected data).

86. 29 C.F.R. §§ 1607.4(A), (C)–(D), 1607.15(A)(2) (2025).

87. 29 C.F.R. §§ 1607.4, 1607.15 (2025); Uniform Guidelines Questions and Answers, *supra* note 70, Questions 85–88.

institutions were forbidden to collect and analyze protected-class data. An interpretation of *SFFA* that forbade it would not leave the framework diminished. It would leave it inoperable.

An algorithmic audit extends that established practice rather than departing from it. An employer may compare the rate at which an automated screener recommends applicants from different racial groups. It may examine whether a video-interview system produces different false-negative rates. It may determine whether a model's confidence scores are less reliable for one population than another. None of those calculations change any applicant's result. They reveal characteristics of the selection process, which is the same thing a validation study has always done.

The distinction can be stated in terms of data flow. In a compliance system, protected-class data move into the audit but not into the individual decision. The operational model produces scores or recommendations without access to the demographic audit file. Auditors later join outcome data with protected-class data to evaluate the system. Access is limited, the purpose is documented, and the demographic information is never returned to the decision maker as a reason to alter a particular result. Described that way, the difference between compliance and classification is not a matter of motive at all. It is a matter of where in the system the data are permitted to go, which is a question about architecture and can be answered by inspecting it.

Local Law 144 rests on the same premise. It prohibits covered employers and employment agencies from using certain automated employment decision tools unless the tool has undergone a recent bias audit, information about the audit is publicly available, and specified notice has been provided.[88] A jurisdiction that requires an employer to measure protected-group effects has necessarily concluded that measuring them is not the same as using them.

88. N.Y.C. Admin. Code § 20-871; *Automated Employment Decision Tools*, N.Y.C. Dep't of Consumer & Worker Prot., https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page [https://perma.cc/2CQV-MBXL] (describing the bias-audit, publication, and notice requirements of Local Law 144).

C. *System-Level Remediation is not Necessarily Classification*

An audit that merely documents a disparity has limited value. Civil rights compliance also requires institutions to determine whether the disparity is justified and whether a less discriminatory system could serve the same legitimate objective. That inquiry will often lead to a change in the model or the process, which is the point at which the classification question becomes pressing.

Suppose an employer's audit reveals that a resume-screening model disproportionately excludes Black applicants. The employer investigates and discovers that the model assigns significant weight to uninterrupted employment history, although the feature adds little to the model's ability to predict job performance. The employer removes that feature and applies the revised model to every applicant. The employer acted because it observed a racial disparity. But it does not follow that the revised system classifies applicants by race. Every applicant is evaluated under the same model, and no applicant's race determines her score. Applying the counterfactual test, altering an applicant's recorded race changes nothing about the output because the model never receives it.

The same analysis applies when an institution revises a target variable. A model trained to predict retention may disadvantage applicants whose prior employment histories do not resemble those of the incumbent workforce. An employer could instead train the model to predict performance on defined work tasks. If the employer makes that change after discovering racial disparity, its motivation is race-conscious. The revised criterion, however, can still be applied uniformly, and it can be defended on its own terms as a more accurate measure of the work. Indeed, the revision moves the model toward what the validation standards required in the first place.[89]

Other system-level responses may include changing a cutoff, eliminating a weak proxy, improving the representativeness of training data, adding an appeal process, requiring review of low-confidence outputs, or selecting another model with similar predictive validity and less adverse impact. None of these changes confers an individual

89. *See* 29 C.F.R. § 1607.5(B)–(C)(2025).

preference. They change the general rule under which all individuals are evaluated.

Title VII already contemplates remediation of this kind. The Uniform Guidelines advise employers to investigate alternative selection procedures when an existing procedure has adverse impact, and federal guidance favors the alternative when it has substantially equal validity and less adverse impact. The EEOC likewise advises employers that they remain responsible for vendor-supplied procedures and should adopt equally effective procedures that produce less adverse impact.[90] The obligation to search for a less discriminatory alternative is also an element of the plaintiff's case under Title VII, which means the law does not merely permit the search but assumes it has occurred.[91]

It would be anomalous to treat compliance with that obligation as racial discrimination merely because group data revealed the need for change. The employer is not choosing an applicant because of race. It is choosing a selection procedure after learning that the original procedure may impose an unjustified racial burden. Civil rights law has required institutions to evaluate their practices in exactly this manner since *Griggs*.[92]

A developer may use protected-class data during training to select among candidate models, or to impose a constraint intended to reduce unequal error rates. The deployed model may then apply a single decision rule to all individuals without receiving protected-class information at all. This practice is more race-conscious than an ordinary post-deployment audit, because demographic information influenced the construction of the model itself rather than merely its evaluation. It is the hardest case for the framework proposed here, and it deserves a direct answer rather than an assimilation to one of the easier categories.

90. 29 C.F.R. § 1607.3(B) (2025); Uniform Guidelines Questions and Answers, *supra* note 70, Questions 30–31; U.S. EQUAL EMP. OPPORTUNITY COMM'N, *Employment Tests and Selection Procedures*, *supra* note 65.

91. 42 U.S.C. § 2000e-2(k)(1)(A)(ii), (C); *see* 29 C.F.R. § 1607.3(B).

92. *Cf.* Griggs v. Duke Power Co. 401 U.S. 424, 43170–36 (1971); Albemarle Paper Co. v. Moody, 422 U.S. 405, 425–36 (1975).

It should nevertheless remain within the compliance category when three conditions are met. First, the developer produces one generally applicable model rather than assigning different models, weights, or thresholds according to an individual's race. Second, the protected-class data are not used to construct a hidden racial classifier or to reproduce race through deliberately selected proxies. Third, the resulting model remains connected to a legitimate and legally defensible objective, rather than being engineered solely to generate a predetermined racial distribution.

Under those conditions, race informs the evaluation of candidate systems but does not govern any individual decision. The practice resembles choosing a validated employment test with less adverse impact. The employer knows that the alternatives have different racial effects, and that knowledge influences its choice, but each applicant takes the same test and is evaluated under the same standard. An employer who selected the less discriminatory of two validated paper examinations in 1975 would not have been thought to classify by race, and nothing about performing the comparison in software changes the analysis.

D. Remediation Becomes Classification when Race Changes Individual Treatment

The line changes when an institution uses race to modify an individual result. Title VII expressly prohibits employers from adjusting test scores, using different cutoff scores, or otherwise altering the results of employment-related tests on the basis of race. That prohibition applies whether the adjustment is performed by a human administrator or embedded in an algorithm.[93]

An employer therefore crosses from system-level remediation to classification when it lowers the hiring threshold for one racial group, applies group-specific model weights, adds or subtracts points according to race, or overrides individual results to achieve a desired racial distribution. Those practices do not merely change the system for everyone. They make a person's race part of the rule determining that person's opportunity.

93. *See* 42 U.S.C. § 2000e-2(l); *supra* note 70 and accompanying text.

New Haven was permitted to know that its promotion examinations produced different racial results. It could collect the data, calculate the disparity, investigate the examinations, hear from experts, and consider alternative procedures. The legal difficulty arose only when the city declined to certify completed results because of their racial distribution, thereby changing the consequences of the examination for identifiable candidates.[94] *Ricci* did not prohibit diagnosis. It regulated a race-conscious response that affected individual employment opportunities, and the response it regulated came after the opportunities had already been earned.

Group-specific calibration is the hardest version of the problem. A developer might conclude that the same raw score carries a different statistical meaning for different racial groups and might adopt separate conversion rules to equalize predictive accuracy. The technical argument for doing so can be entirely sincere. A body of work establishes that a risk score cannot generally satisfy calibration within groups and equal error rates across groups at the same time, except in circumstances that rarely occur, so a developer confronting real data must choose which property to sacrifice.[95] But the mechanism is a lookup by racial category: The applicant's group determines which conversion applies, and two applicants with identical raw scores receive different treated scores. Race would determine which scoring rule applies to a given individual. That is classification and should be analyzed as such. Technical sophistication does not eliminate the constitutional or statutory concern created by differential treatment, and a statistical impossibility result does not supply a defense that the statute withholds.[96]

94. Ricci v. DeStefano, 557 U.S. 557, 574–75, 592–93 (2009).

95. *See* Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, in 8TH INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE 43:1, 43:3 (2017), https://doi.org/10.4230/LIPIcs.ITCS.2017.43 [https://perma.cc/96DT-D4ZE]; Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153 (2017).

96. *See supra* note 28 and accompanying text. (Section 703(l) bars the use of "different cutoff scores" on the basis of race without regard to the statistical rationale offered for the difference).

The same is true of automated post-processing designed to impose a racial quota. A system that re-ranks applicants until a desired racial distribution is reached uses race as an allocation mechanism even if the original model did not. Calling the process "bias correction" does not answer the legal question. Courts must examine how the correction actually operates and determine whether it changes individual treatment because of race.

This boundary does not make remediation impossible. It directs institutions toward system-level changes before individual racial adjustments. An institution can reconsider what it is predicting, remove poorly justified inputs, strengthen validation, adopt a less discriminatory alternative, improve process safeguards, or replace automation with a more individualized assessment. Each of those changes may be undertaken with full knowledge of racial consequences while still applying a common decision rule to everyone. The ordering is not merely prudential. It reflects the difference between fixing a practice and adjusting a result.

E. A Protected Audit Safe Harbor

Courts and agencies should formalize the distinction through a protected audit safe harbor. The safe harbor would establish a rebuttable presumption that the collection and use of protected-class data for algorithmic compliance do not constitute unlawful racial classification when specified safeguards are satisfied.[97]

The safe harbor would protect an institution's authority to conduct the audit. It would not immunize the system being audited. An employer could not invoke it to avoid liability for a hiring tool that violates Title VII. Nor would the safe harbor establish that a particular statistical method is adequate or that an identified disparity is legally justified. It answers a narrower antecedent question: whether the institution violated the law merely by collecting and analyzing protected-class information in order to find out whether the tool discriminates.

97. The safe harbor proposed here concerns only the legality of using protected-class data for civil rights compliance. It would not displace otherwise applicable privacy, cybersecurity, records-retention, or data-governance duties.

First, the institution must identify a bona fide compliance purpose. Protected-class data may be used for auditing, validation, monitoring, regulatory reporting, investigation, or evaluation of less discriminatory alternatives. A generalized desire to balance outcomes would not suffice. The institution should be able to connect the analysis to an applicable civil rights obligation, a recognized risk assessment, or a documented effort to determine whether the system operates fairly and lawfully.[98]

Second, the institution must maintain operational separation. Protected-class data should not be available to the model or to the decision maker when an individual decision is made, unless another law independently authorizes that use. Demographic information should ordinarily be stored separately, access should be restricted, and each use should be documented. Separation is what makes the classification-compliance distinction observable rather than rhetorical, because it produces a record of where the data went.

Third, the audit must examine the actual decision process. A vendor should not be able to satisfy the safe harbor by testing a model under artificial conditions while ignoring the threshold, data, population, or workflow through which the deployer uses it. A model that shows no disparity on the vendor's benchmark data may show a substantial one at the cutoff a particular employer selected, applied to the applicant pool that employer actually draws. Depending on the context, the audit should therefore examine selection rates, error rates, calibration, missing data, overrides, appeals, and the interaction between the model and human decision makers. The Uniform Guidelines' distinction between a component procedure and the total selection process supplies a workable framework for that inquiry.[99]

Fourth, any remediation within the safe harbor must initially operate at the system level. The institution may adopt a common model with less adverse impact, revise a target, remove an unjustified variable, change a generally applicable threshold, or add review

98. *See* Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems, *supra* note 19; U.S. Equal Emp. Opportunity Comm'n, Select Issues, *supra* note 37.

99. 29 C.F.R. § 1607.16(Q) (2025); Uniform Guidelines Questions and Answers, *supra* note 70, Questions 11–13.

procedures available on the same terms to everyone. When race determines which rule or threshold applies to a person, the safe harbor ends and ordinary constitutional and statutory scrutiny begins.

Fifth, the institution must monitor the revised system rather than treating one audit as permanent validation. Models change when their inputs, users, applicant populations, or institutional contexts change. A tool that produces little disparity in one setting may operate quite differently in another, and a model retrained on data generated by its own prior recommendations may drift in ways no one intended. Monitoring should be proportionate to the stakes and to the system's capacity to change.

These safeguards create a rebuttable presumption rather than categorical immunity. A challenger could overcome the presumption by showing that protected-class data entered individual decisions, that the audit was a pretext for racial allocation, that the institution deliberately reconstructed race through proxies, or that the purportedly uniform redesign functioned as a mechanism for assigning outcomes by race.

The safe harbor also should not create a broad privilege against discovery. Shielding all audit results would deprive plaintiffs and regulators of some of the most probative evidence about how a system operates and would do so precisely in the cases where that evidence is hardest to obtain elsewhere.[100] Institutions may need carefully defined protections for preliminary analyses, cybersecurity information, trade secrets, or personal data. But the act of auditing should not become a means of concealing discrimination. The safe harbor protects the legality of good-faith measurement. It does not suppress what the measurement reveals.

F. Race-Conscious Purpose is not Race-Based Treatment

The strongest objection to this framework is that system-level remediation may still be undertaken for an explicitly racial purpose. An employer selects one model over another because the audit shows that the first disproportionately excludes Black applicants. A lender changes a threshold because it has observed a racial disparity. A public

100. For discussion of the informational asymmetry created by data-driven employment systems, *see* Kim, *supra* note 56, at 905–10.

agency revises a risk tool because the existing tool concentrates errors in one racial group. Race plainly influenced each institutional decision, and pretending otherwise would be dishonest.

But equal protection doctrine does not ordinarily subject every decision made with awareness of racial effects to strict scrutiny. The more precise inquiry asks whether the government classified individuals by race or subordinated its ordinary decision criteria to race in a manner that functions as a classification. Justice Kennedy's *Parents Involved* concurrence recognized that government may track outcomes by race and adopt general measures informed by racial conditions without assigning any particular person a benefit or burden according to race.[101]

A contrary rule would make knowledge of discrimination disabling. Once an institution learned that a facially neutral practice produced an unjustified racial disparity, its awareness of the disparity would prevent it from adopting a better uniform practice without risking a claim that it had acted "because of race." Institutions would then have a legal incentive to avoid measurement, preserve ignorance, and continue using the original system, which is the opposite of what the civil rights statutes were written to encourage. The rule would also be self-defeating on its own terms, since an institution that never measures cannot discover the classifications it may already be running.

The Constitution does not require that result. A single selection procedure chosen because it is valid and produces less unjustified adverse impact does not treat any applicant differently because of race. It treats every applicant under a revised rule. Race may explain why the institution investigated the original practice and why it compared alternatives. It does not determine which rule applies to which person.

The classification-compliance distinction must therefore be operational rather than psychological. Courts should not ask only whether officials thought about race. They should ask where protected-class information entered the process, whether it altered individual treatment, whether the revised system applies a common rule, and whether the institution can defend that rule by reference to

101. Parents Involved in Cmty. Sch's. v. Seatle Sch. Dist. No. 1, 551 U.S. 701, 788–89 (2007) (Kennedy, J., concurring in part and concurring in the judgment); *see supra* note 78 and accompanying text.

the legitimate objective the system is supposed to serve. Those questions have answers that do not depend on reconstructing anyone's state of mind, which is an advantage in a field where the decision maker is often a piece of software.

A legal regime capable of making those distinctions can preserve *SFFA's* prohibition against racial allocation without disabling disparate-impact enforcement. It can permit institutions to see racial inequality, investigate its causes, and redesign flawed systems while maintaining a firm boundary against using race to distribute individual opportunities. The alternative collapses observation into discrimination and leaves facially neutral algorithms uniquely protected from scrutiny. Under that approach, the system may reconstruct race through hundreds of variables, but the institution charged with complying with civil rights law may not name what the system has done.

VII. Civil Rights Law Must Be Able to See what AI Hides

AI changes the grammar of discrimination. Older civil rights disputes turned on visible acts of exclusion: a rule, a policy, a decision maker, a classification. AI systems are more diffuse. They operate through prediction, correlation, optimization, and ranking; and those operations do not map cleanly onto the categories the statutes were written around. That does not make them immune from civil rights law. It makes civil rights law harder to apply.

If courts focus too narrowly on whether race was explicitly used, they will miss the systems that reconstruct race through proxies. If institutions are told that race-conscious auditing is legally risky, many will simply choose not to know, and they will be acting rationally when they do. That is not colorblindness as equality. A lender that does not measure racialized effects is not less discriminatory than one that does. An employer that avoids auditing its AI tools has not achieved equality by avoiding evidence. It has achieved deniability, which the civil rights statutes were written to take away.

The post-*SFFA* challenge is therefore to preserve a narrow but essential form of race-consciousness. Law can forbid racial allocation while permitting racial measurement for compliance. It can reject racial stereotyping while requiring institutions to test whether their

tools produce racialized harm. It can limit race-based remedies while allowing race-conscious diagnosis. Each of those pairs describes a line the law already knows how to draw in other settings, and none of them requires a new constitutional category.

Algorithmic systems can make discrimination look like prediction, and prediction look like merit. A model does not have to say "race" to reproduce racial hierarchy. It can say "risk," "fit," "default," "engagement," "safety," or "merit." Those labels are not self-validating. They are choices about what to measure and what to reward. Civil rights law must therefore ask what they are built from, how they operate, and who bears their errors. *SFFA* should not be read to require blindness to those questions. The danger of algorithmic governance is not that institutions will use race explicitly. The danger is that they will not need to. Race-conscious compliance is not an exception to civil rights law. It is a condition of making civil rights law work.