Stanford Law Review Online

Volume 79 September 2026

ESSAY

Classifying and Countering AI Proxy Discrimination

Keith Swisher*

Introduction

> *We have passed laws prohibiting discrimination in education, in employment, in housing; but these laws alone cannot overcome the heritage of . . . poverty and degradation and pain.*
>
> —Robert F. Kennedy[1]
>
> *The law, like the traveler, must be ready for the morrow. It must have a principle of growth.*
>
> —Benjamin N. Cardozo[2]

Derek Mobley applied to over one hundred jobs through employers using Workday's AI screening platform. He was rejected for every job, sometimes through emails sent at 2:00 AM when no human was reviewing applications. Mobley is Black, over forty, and lives with a disability. He alleged that Workday's algorithm systematically filtered out applicants like him. But when he sued, the court dismissed his intentional discrimination claims: An algorithm, after all, has no intent. Although discrimination seemed apparent in the allegations, the law's protection was unclear and uncertain.[3] Likewise,

* Professor, University of Arizona James E. Rogers College of Law. I owe many thanks to Renee Knake Jefferson, Marc Miller, and Justin Pidot for excellent comments on earlier drafts and to Hannah L. Dahleen and Kaidi Zhang of the *Stanford Law Review Online* for substantial improvements. Any errors are mine alone.

1. Senator Robert F. Kennedy, Day of Affirmation Address, University of Capetown (June 6, 1966), https://perma.cc/XY4U-STTA.

2. BENJAMIN N. CARDOZO, THE GROWTH OF THE LAW 20 (1924).

3. *See* Mobley v. Workday, Inc., 740 F. Supp. 3d 796, 812-13 (N.D. Cal. 2024); Mobley v. Workday, Inc., No. 23-cv-00770, 2025 WL 1424347, at *1 (N.D. Cal. May 16, 2025) (certifying collective action); *cf.* Washington v. Davis, 426 U.S. 229, 239-42 (1976) (requiring proof of discriminatory purpose under the Equal Protection Clause). Mobley's disparate impact claims are pending at the time of this Essay's publication.

when the U.S. Department of Health and Human Services (HHS) in 2024 finally extended nondiscrimination principles to AI-driven, decision-support tools in patientcare, it was responding to a problem the law had missed: a widely used health-care algorithm had been assigning Black patients lower risk scores by treating prior spending as a proxy for need. As these recent examples indicate, algorithms can discriminate in ways that old doctrine struggles to catch.[4]

Yet American civil rights law is organized around protected classes. Title VII of the Civil Rights Act forbids discrimination on the basis of race, color, religion, sex, and national origin.[5] The Fair Housing Act prohibits discrimination in housing on similar grounds.[6] The Americans with Disabilities Act protects individuals with disabilities.[7] The Voting Rights Act safeguards the franchise against racial discrimination.[8] In each statute, the structure of protection is the same: the law identifies a characteristic (race, sex, disability) and prohibits decisions that disadvantage individuals on the basis of that characteristic. This protected-class framework has anchored civil rights enforcement for over sixty years.

Algorithmic decisionmaking systems outmode this framework.[9] Machine learning models typically do not sort individuals into the law's protected classes; they generate their own classifications, constructed from hundreds or

4. S*ee* Nondiscrimination in Health Programs and Activities, 89 Fed. Reg. 37522, 37646-50 (May 6, 2024) (codified at 45 C.F.R. § 92.210) (addressing discrimination through the use of patient care decision support tools but declining to reach proxy discrimination explicitly); Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 447, 449-50 (2019).

5. S*ee* Civil Rights Act of 1964, Pub. L. No. 88-352, § 703(a), 78 Stat. 254, 255 (codified as amended at 42 U.S.C. § 2000e-2(a)).

6. Fair Housing Act of 1968, Pub. L. No. 117-103, § 804(b), 82 Stat. 73, 83 (codified as amended at 42 U.S.C. § 3604(f)(3)(A)-(B)).

7. Americans with Disabilities Act of 1990, Pub. L. No. 101-336, 104 Stat. 327 (codified as amended in scattered sections of the U.S. Code).

8. Voting Rights Act of 1965, Pub. L. No. 89-110, § 2, 79 Stat. 437, 437 (codified as amended at 52 U.S.C. § 10301).

9. This Essay addresses "algorithmic decisionmaking systems" and "AI" in broad terms, including machine learning and Large Language Models (LLMs). The discussion is meant to be inclusive of any current or future technology classifying individuals to inform or make decisions in civil rights domains and potentially using proxy discrimination in its classifications. *See generally* Kyra Wilson & Aylin Caliskan, *Gender, Race, and Intersectional Bias in Resume Screening Via Language Model Retrieval*, 7 PROCS. AAAI/ACM CONF. AI, ETHICS, & SOC'Y 1578, 1578-88 (2024), https://perma.cc/6H6K-DFN8 (finding bias in language models); Haozhe An, Christabel Acquaye, Colin Kai Wang, Zongxia Li & Rachel Rudinger, *Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender?*, 2 PROCS. 62ND ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 386 (2024), https://perma.cc/6ZY8-9YKV (same); Rishi Bommasani, Sarah H. Bana, Kathleen A. Creel, Dan Jurafsky & Percy Liang, *Algorithmic Monocultures in Hiring*, 2026 PROCS. ACM CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 6351 (2026), https://perma.cc/W48S-VEM8 (finding racial disparities when using job applicant screening algorithms).

thousands of data features that bear no obvious relationship to race, sex, or other prohibited grounds.[10] Yet these algorithmically generated categories can function as precise proxies for protected characteristics, sorting individuals into groups that map onto race or sex with high fidelity while never naming either.[11] The result is a form of discrimination that operates beneath the conceptual threshold of existing civil rights law: the algorithm discriminates by proxy, and the law's focus on named characteristics leaves it largely unable to respond.

Proxy discrimination is not new, but this Essay argues that proxy discrimination through algorithmic classification represents a qualitatively new challenge to antidiscrimination law. Incremental doctrinal adjustments to disparate impact or disparate treatment frameworks cannot fix the problem; instead, novel approaches are necessary. Part I explains the limitations of the status quo.

Then, drawing on the philosophy of science, information theory, and behavioral economics, I propose three interdisciplinary reforms. Part II provides an epistemological test for proxy discrimination grounded in the concept of mutual information. Part III offers classification impact assessments modeled in part on environmental impact and the philosophy of science's treatment of classificatory systems. Finally, Part IV proposes a fairness-by-design rule inspired by behavioral economics' concept of choice architecture. Individually or collectively, these reforms will help to detect and mitigate modern proxy discrimination.

10. S*ee* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677-80 (2016).

11. S*ee* Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1260-65 (2020). For two recent case examples, see Complaint, Harper v. Sirius XM Radio, LLC, No. 25-cv-12403 (E.D. Mich. Aug. 4, 2025), ECF No. 1 (alleging that the defendant used iCIMS AI/ML tools to screen, rank, or filter applicants using race-correlated data points—including educational institutions, employment history, and zip codes—and disproportionately excluded African American applicants, in violation of Title VII and 42 U.S.C. § 1981); Huskey v. State Farm Fire & Cas. Co., No. 22 C 7014, 2023 WL 5848164, at *1-2, *9-11 (N.D. Ill. Sept. 11, 2023) (denying dismissal of a disparate-impact claim under § 3604(b) of the Fair Housing Act based on allegations that algorithmic claims-processing tools produced statistically significant racial disparities and dismissing without prejudice claims under §§ 3604(a) and 3605); *see also* Michael R. Greco & Karen L. Odash, *Another Employer Faces AI Hiring Bias Lawsuit: 10 Actions You Can Take to Prevent AI Litigation*, FISHER PHILLIPS (Aug. 15, 2025), https://perma.cc/CZ7G-8P8G (discussing *Harper*); *State Farm Algorithm Bias Lawsuit*, SANFORD HEISLER SHARP MCKNIGHT, LLP, https://perma.cc/4PWH-63VV (archived Aug. 13, 2026) (discussing *Huskey*).

I. Classification and the Structure of Civil Rights Law

A. The Protected-Class Paradigm

The central organizing principle of American antidiscrimination law is the protected class. Under Title VII, an employer may not "fail or refuse to hire or . . . discharge any individual, or otherwise . . . discriminate against any individual . . . because of such individual's race, color, religion, sex, or national origin."[12] Under the Fair Housing Act, it is unlawful to "refuse to sell or rent . . . or to refuse to negotiate for the sale or rental of, or otherwise make unavailable or deny, a dwelling to any person because of race, color, religion, sex, familial status, or national origin."[13] Under the Equal Protection Clause, classifications on the basis of race are subject to strict scrutiny, while classifications on the basis of sex receive intermediate scrutiny.[14]

This framework assumes that discrimination operates through classifications recognizable to the law. An employer who refuses to hire women, a landlord who excludes Black tenants, and an agency that denies benefits to people with disabilities are all paradigm cases. The law identifies the protected characteristic, the decisionmaker's intentional reliance on that characteristic, and the resulting harm. The doctrinal apparatus of disparate treatment and disparate impact rests on this underlying classificatory logic.[15]

Even disparate impact doctrine, which extends protection beyond intentional discrimination, is tethered to protected classes. A plaintiff generally must show that a facially neutral practice produces a disproportionate adverse effect on a protected group—defined by the statute's enumerated characteristics.[16] The doctrine does not ask whether the practice creates disadvantage along some other axis of classification; it instead typically asks whether the practice produces statistical disparities measured against the statute's named categories. The protected class, in short, is both the gateway to legal protection and the metric by which discrimination is assessed.

B. How Algorithms Classify

Machine learning systems generate their own classifications. Rather than relying on the law's protected categories, a machine learning model constructs a decision boundary—a mathematical function that sorts inputs into categories (e.g., approved/denied, high-risk/low-risk, hired/flagged)—by identifying

12. 42 U.S.C. § 2000e-2(a).

13. 42 U.S.C. § 3604(f)(3)(A)-(B).

14. S*ee, e.g.*, Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 227 (1995) (concluding that racial classifications are subject to strict scrutiny); Craig v. Boren, 429 U.S. 190, 197 (1976) (concluding that sex-based classifications are subject to intermediate scrutiny).

15. S*ee* Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

16. S*ee* 42 U.S.C. § 2000e-2(k)(1)(A)(i).

patterns in training data.[17] The model's classifications are emergent properties of its training process: the categories it creates are not designed by a human decision-maker but are "learned" by the model from the statistical structure of the data.

These learned classifications may bear little surface resemblance to the law's protected categories. An algorithmic hiring tool almost never classifies applicants as "Black" or "White"; it classifies them as higher or lower scoring based on features such as educational background, commute distance, word choice in application materials, employment gaps, and behavioral assessment results.[18] A credit scoring model does not classify applicants by race; it classifies them by payment history, debt-to-income ratio, and dozens of other variables.[19] A criminal risk assessment generally does not classify defendants by race; it classifies them by charge type, criminal history, age at first arrest, and social stability indicators (among others).[20]

Yet these algorithmically constructed categories can reproduce the effects of racial classification with striking precision. Given racial inequality throughout society, features such as zip code, educational attainment, arrest history, and employment pattern can be strongly correlated with race.[21] A prediction-optimizing algorithm will identify and exploit these correlations, because in a racially unequal society, race and its proxies are predictive of the outcomes the model is trained to forecast.[22] Prince and Schwarcz, for example, have documented this phenomenon extensively, showing that algorithmic systems can reconstruct protected-class membership from facially neutral data with high accuracy.[23]

---

17. *See, e.g.*, Barocas & Selbst, *supra* note 10, at 677-80 (explaining this process in the context of data mining).

18. S*ee, e.g.*, Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 874-80, 886-92 (2017) (discussing this phenomenon and classification bias).

19. S*ee, e.g.*, Robert Bartlett, Adair Morse, Richard Stanton & Nancy Wallace, *Consumer-Lending Discrimination in the FinTech Era*, 143 J. FIN. ECON. 30, 31 (2022) (discussing racial disparities in algorithmic-aided consumer lending outcomes).

20. S*ee* Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. 439, 450-54 (2020).

21. S*ee, e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2218 (2019) ("In a racially stratified world, any method of prediction will project the inequalities of the past into the future.").

22. S*ee id.* at 2224-26 (noting that given widespread racial inequality, race-correlated features will inevitably be predictive of outcomes); Ngozi Okidegbe, *Discredited Data*, 107 CORNELL L. REV. 2007, 2012-18 (2022).

23. *See* Prince & Schwarcz, *supra* note 11, at 1260-65, 1273-76; *see also* Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 HASTINGS L.J. 1389, 1413 (2019) ("The proxies that algorithmic processes might identify, or even the fact of whether an algorithm will identify a proxy at all, is difficult—and sometimes impossible—to predict.").

The result is a classification system that produces discriminatory outcomes without directly engaging the law's protected-class categories. The algorithm classifies, but it classifies on its own terms, generating associations and categories that the existing legal framework does not recognize and cannot easily regulate.

II. The Epistemological Problem: How Do We Know When a Proxy Is a Proxy?

The challenge of proxy discrimination is not merely technical; it is epistemological. How do we determine whether an algorithmically generated classification constitutes a proxy for a protected characteristic? Existing law provides no satisfactory answer. Under the intent doctrine of *Washington v. Davis*, a facially neutral practice is permissible absent evidence that the decisionmaker selected it "because of" its discriminatory effects.[24] Under *Personnel Administrator of Massachusetts v. Feeney*, even foreseeable discriminatory effects are insufficient to establish intent if the decisionmaker acted "in spite of," rather than "because of," those effects.[25] These doctrines are essentially useless in the algorithmic context because the "decisionmaker" is a statistical optimization process that has no intent at all.[26]

Although disparate impact doctrine generally fares better, it still founders on the proxy problem. For example, a plaintiff may establish a prima facie case by showing that a practice produces statistically disproportionate effects on a protected group.[27] But which practice? If the algorithm relies on one hundred features, some of which are correlated with race and some not, the plaintiff faces an identification problem: there may be no single feature that "causes" the disparity, because the disparity emerges from the holistic interaction of many features.[28] Thus, this places a difficult and unrealistic burden on plaintiffs. Furthermore, this makes the doctrine imprecise; it seeks discriminatory practices but lacks the means to identify them. Finally, even if a single feature can be identified—say, zip code—the employer can argue that zip code is not a proxy for race but a legitimate predictor of the relevant outcome (e.g., commute time or local labor market conditions).

---

24. 426 U.S. 229, 239-42 (1976).

25. 442 U.S. 256, 279 (1979) (requiring that the course of action be "at least in part 'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group").

26. S*ee* Note, *Beyond Intent: Establishing Discriminatory Purpose in Algorithmic Risk Assessment*, 134 HARV. L. REV. 1760, 1764-69 (2021). To be sure, those who develop or implement the system may have discriminatory intent, but the algorithm or AI itself is incapable of intent.

27. *See* Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

28. *See* Barocas & Selbst, *supra* note 10, at 677-80.

This Part addresses the proxy problem by (A) exploring and exposing the impactful nature of certain classifications and (B) proposing a mutual information test to detect proxy discrimination warranting increased legal scrutiny.

A. Insights from the Philosophy of Science: Classifications as Interactive Kinds

The philosophy of science provides a productive framework for thinking about algorithmic classifications and their relationship to the law's protected categories. Ian Hacking has distinguished between "natural kinds" and "interactive kinds." Natural kinds are categories that exist independently of human classification, such as chemical elements or Homo sapiens. Interactive kinds, in contrast, are categories whose members are affected by the very act of classification.[29] The key insight here is that social classifications do not merely describe a pre-existing reality; they constitute that reality. People who are classified as "criminals," "welfare recipients," or "high-risk borrowers" are shaped by those classifications: the classifications affect how they are treated, what opportunities are available to them, and even how they understand themselves.[30]

Algorithmic classifications are very much interactive kinds. When a risk assessment labels a defendant "high-risk," that classification does not merely describe a pre-existing reality; it creates consequences, such as pretrial detention or harsher sentencing, which in turn make future negative outcomes more likely (loss of employment, family disruption, exposure to jail or prison environments that increase recidivism).[31] When a credit-scoring algorithm classifies an applicant as "subprime" or the like, the classification produces increased interest rates and decreased access to financial services, which in turn make financial distress more likely.[32] The algorithm's classifications, in other words, are partly self-fulfilling: they produce the very outcomes they purport merely to classify or predict.

Bowker and Star's work on the politics of classification systems generally reinforces this point. They argue that all classification systems embed normative choices about which differences matter and which do not—choices

29. S*ee* IAN HACKING, THE SOCIAL CONSTRUCTION OF WHAT? 103-07 (1999) (discussing natural and interactive kinds in philosophy, natural science, and social science). Natural kinds are also known as indifferent kinds.

30. S*ee id.*

31. S*ee, e.g.*, CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 1-11, 84-100 (2016) (discussing pernicious feedback loops in employment, credit, and criminal justice).

32. *See, e.g.*, Bartlett et al., *supra* note 19, at 55 (discussing algorithmic harm in consumer lending).

typically invisible to the system's users.[33] Algorithmic classification systems embed these choices in mathematical or statistical form, but the selection of features, the definition of the target variable, and the optimization of the model all involve normative decisions about what counts as a relevant distinction. When these distinctions track protected characteristics (including indirectly through proxies), the classification system is doing the work of discrimination without using the law's vocabulary of discrimination.

This philosophical analysis suggests that the law's focus on whether a decisionmaker "used" a protected characteristic is the wrong question. The right question is instead whether the classification system produces categories informationally entangled with protected-class membership—in other words, whether knowing an individual's algorithmic classification tells you, with high probability, the individual's race, sex, age, disability, or other protected characteristic. If it does, and if this leads to negative outcomes, the classification is functionally discriminatory regardless of whether anyone or anything "intended" it to be or explicitly "used" a protected characteristic. The law thus should employ a test capable of identifying AI proxy discrimination, which is proposed below. The need is even more compelling when we consider the self-fulfilling qualities and normative nature of classification highlighted above.

B. An Information-Theoretic Test for Proxy Discrimination

Information theory offers a precise, quantitative framework for operationalizing the insight above. The concept of mutual information, developed by Claude Shannon and formalized in modern information theory, measures the degree of statistical dependence between two random variables.[34] If the mutual information between an algorithmic classification and a protected characteristic is high, then the classification is informationally redundant with the protected characteristic: knowing one tells you a great deal about the other.

This Essay proposes that lawmakers should update civil rights law to include a mutual information threshold as a formal test for proxy discrimination. Specifically, an algorithmic classification should be presumptively discriminatory if the mutual information between the classification and any protected characteristic exceeds a defined threshold. This threshold would function analogously to the Four-Fifths Rule in the Uniform

33. *See* GEOFFREY C. BOWKER & SUSAN LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES 3-16 (1999).

34. *See* THOMAS M. COVER & JOY A. THOMAS, ELEMENTS OF INFORMATION THEORY 12, 18-23 (1991) (noting that mutual information is in part a "measure of the amount of information that one random variable contains about another random variable").

Guidelines on Employee Selection Procedures,[35] which establishes a quantitative benchmark for adverse impact. In particular, if an employment selection practice results in a protected group being selected at a rate of less than four-fifths (or 80 percent) of the highest selected group, this constitutes evidence of disparate impact. But whereas the Four-Fifths Rule measures the ratio of selection rates between groups, the mutual information test measures the degree to which the classification itself is informationally entangled with protected-class membership.

This approach has several advantages. First, it avoids the identification problem that plagues disparate impact doctrine. The test does not require the plaintiff to isolate a single feature or practice; it evaluates the classification system as a whole.[36] Second, it captures proxy discrimination directly. Rather than asking whether a particular feature is "really" a proxy for race—a question that often degenerates into contested empirical and normative judgments—the test asks whether the algorithm's output is statistically entangled with race.[37] Third, the test is computationally manageable: mutual information can ordinarily be calculated from the same data used to train and validate the algorithm, requiring no additional data collection.

Of course, a threshold must be chosen, and any choice will be somewhat arbitrary. But this is no different from existing antidiscrimination thresholds. The Four-Fifths Rule's 80-percent benchmark is not derived from first principles; it is a policy choice reflecting a judgment about the level of disparity that warrants scrutiny.[38] A mutual information threshold would reflect an analogous judgment about the degree of informational entanglement between algorithmic classifications and protected characteristics triggering legal concern. The threshold could be calibrated over time as agencies and courts gain experience with the framework.

---

35. S*ee* 29 C.F.R. § 1607.4(D) (2015); Keith Swisher, *Algorithmic Judicial Ethics*, 2024 WIS. L. REV. 1289, 1324-27; Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1119-20 (2019).

36. *See* Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 657 (1989) (requiring identification of specific employment practice at issue); Barocas & Selbst, *supra* note 10, at 694-712 (discussing the holistic nature of machine learning models and the difficulty of isolating individual features).

37. C*f.* Matthew U. Scherer, Allan G. King & Marko J. Mrkonich, *Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms*, 71 S.C. L. REV. 449, 508-09 (2019) (applying statistical independence concepts to the employment discrimination context).

38. *See* 29 C.F.R. § 1607.4(D). *See also* Frank Fagan, *Proxy Discrimination After* Students for Fair Admissions, 8 J.L. & TECH. TEX. 91, 109-22 (2025) (proposing a comparative "proxy power" standard under which a decision tool is narrowly tailored when it exhibits the weakest total proxy power relative to available alternatives measured through regression-based power and arguing that lawmakers should place caps on permissible proxy power).

For a simple illustration of this approach, consider an AI hiring model sorting 1,000 applicants into "recommend" and "do not recommend" folders. Of 500 white applicants, it recommends 400; of 500 Black applicants, it recommends only 100. The AI tool did not explicitly use race in sorting the applicants, but importantly for proxy discrimination, the model's output is highly informative about race: if one knows that an applicant was "recommended," one can predict with high confidence that the applicant is white; if one knows that the applicant was not recommended, one can predict with high confidence that the applicant is Black.

In information-theoretic terms, the mutual information between the model's output and race is materially above zero (about 0.278 bits in this hypothetical), meaning the output is carrying race information rather than merely correlating with race in some loose sense.[39] (In this binary example, 0 bits would mean no informational entanglement with race while 1 bit would mean complete entanglement.[40]) This entanglement is broader than an ordinary correlation coefficient: whereas correlation indexes the strength of linear association between two variables, the mutual information test quantifies the amount of protected-class information the classification carries, whatever the functional form of the dependence.[41] If a jurisdiction set a mutual-information ceiling at, say, 0.1 bits (or a normalized ceiling at 10 percent), the model's usage would presumptively qualify as proxy discrimination even if no single input variable, viewed in isolation, could be labeled the racial proxy.

Thus, as the above application shows, a mutual information test can capture what old approaches often miss. It can handle and identify proxy

39. S*ee* COVER & THOMAS, *supra* note 34, at 18-23 (defining mutual information as the relative entropy between the joint distribution and the product of the marginal distributions). The basic formula is $I(X;Y) = \Sigma_x \Sigma_y p(x,y) \log_2[p(x,y)/(p(x)p(y))]$, where the sums range over all possible values of X and Y, X denotes race, and Y denotes the model's output. In the hypothetical, P(white, recommend) = 0.40, P(white, not recommend) = 0.10, P(Black, recommend) = 0.10, and P(Black, not recommend) = 0.40. Each marginal probability equals 0.50, so $p(x)p(y) = 0.25$ for every cell. Thus, $I(X;Y) = 2[0.40 \cdot \log_2(0.40/0.25)] + 2[0.10 \cdot \log_2(0.10/0.25)] = 2(0.271229) + 2(-0.132193) \approx 0.278$ bits.

40. Because mutual information is bounded by the variables' entropies, its ceiling shifts with the number of categories. Thus, a nonbinary application could raise the ceiling above one. Normalizing by the entropy of the protected characteristic would place models on a common zero-to-one scale, permitting a single threshold of, say, 10 percent to operate consistently across classifiers and demographics.

41. A high correlation coefficient and high mutual information are related but distinct. The correlation coefficient measures the strength and direction of linear association between numeric variables and, outside special cases, can equal zero even when the variables are strongly dependent; mutual information captures dependence of any functional form, is well-suited for the categorical variables predominating the antidiscrimination context, and equals zero only if the variables are statistically independent. *See, e.g.*, COVER & THOMAS, *supra* note 34, at 18; *cf.* Scherer et al., *supra* note 37, at 508-09 (applying statistical-independence concepts to the employment-discrimination context).

discrimination in a world of machine learning. More specifications and refinements (by civil rights domain, by redundance tolerance, and so on) would quickly operationalize the test. The test cannot fully succeed in isolation, however, as the next two Parts suggest.

III. Classification Impact Assessments

The mutual information test provides a tool for identifying proxy discrimination after an algorithmic system has been deployed. But a comprehensive response to algorithmic proxy discrimination also requires proactive assessment of classification systems before deployment. This Part proposes a regime of Classification Impact Assessments (CIAs), modeled on environmental impact assessments and informed by the philosophy of science's treatment of classificatory systems.

A. The Case for Proactive Assessment

Environmental law regulators recognized decades ago that certain government actions carry risks of significant harm best assessed and addressed before the action is taken, rather than through after-the-fact litigation. As a key example, Congress passed the National Environmental Policy Act (NEPA), which requires federal agencies to prepare environmental impact statements for major actions "significantly affecting the quality of the human environment."[42] Moreover, states have "little NEPA" analogs reaching state actions, and private actions are often pulled into early regulatory review because they receive or want to receive federal or state approval, permitting, or licensing.[43] Professor Danielle Citron has argued persuasively that automated decisionmaking systems should be subject to analogous pre-deployment review, which she terms "technological due process."[44]

This Essay extends this insight by proposing a new form of impact assessments focusing specifically on the classificatory dimension of algorithmic systems—that is, on the categories the system creates and their relationship to the law's protected classes.[45] Thus, this CIA proposal is narrower than the increasingly common call for AI risk assessments or impact

42. S*ee* National Environmental Policy Act of 1969, Pub. L. No. 91-190, § 102(2)(C), 83 Stat. 852, 853-54 (1970) (codified as amended at 42 U.S.C. §§ 4321-4347).

43. S*ee, e.g.*, CAL. PUB. RES. CODE §§ 21065, 21100(a), 21151(a) (West 2026); N.Y. ENV'T CONSERV. LAW § 8-0109(2)-(4) (McKinney 2006); N.Y. COMP. CODES R. & REGS. tit. 6, § 617.1(c) (2026) (requiring environmental impact assessments before project approvals).

44. S*ee* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1252-58 (2008).

45. *See id.* at 1308-313 (arguing for transparency, testing, public participation, and formal or informal rulemaking procedures in automated agency decisions).

assessments generally.[46] A CIA would essentially require the deploying entity to answer three questions before the system is used in any high-stakes decisionmaking context: What categories does this system create? What is the relationship between those categories and protected-class membership (measured, for example, by mutual information)? Finally, what are the foreseeable consequences of those categories for members of protected groups?

The last question is particularly important in light of the interactive-kinds analysis described above. Algorithmic classifications are not inert or indifferent descriptions; they produce consequences that feed back into future data.[47] A CIA should therefore assess not only the static relationship between algorithmic categories and protected classes but also the dynamic effects of the classification over time. The questions thus start with the immediate impact on hiring, housing, liberty, or otherwise for protected classes but then reach beyond: Will the classification produce feedback loops that amplify existing disparities? Will it create new categories of disadvantage that the law does not currently or fully recognize? These are questions that an epistemologically informed assessment should address.

A plausible counterargument would point to the difficulty of fully testing certain models before deployment. For example, in assessing LLM-based systems, the CIA would need to include sampling of the current or reasonably anticipated prompts, as the prompts can vary and can inject or spur discrimination.[48] But robust pre-deployment testing is certainly attainable. As key examples, the developers could run and document simulations (and synthetic or hypothetical data could be used where needed), or the jurisdiction could implement a limited rollout or regulatory sandbox to assess the system's likely usage (e.g., prompts and variations) and impact before widespread deployment.

B. Institutional Design

To capture expertise and efficiency, the government agency with jurisdiction over the relevant domain should presumably administer the CIA: the Equal Employment Opportunity Commission (EEOC) for employment algorithms,[49] the U.S. Department of Housing and Urban Development (HUD)

46. S*ee, e.g.*, Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347, 1380 (2023) (discussing AI impact assessments).

47. *See, e.g.*, O'NEIL, *supra* note 31, at 86-97 (discussing pernicious feedback loops in algorithmic systems).

48. In assessing LLM-based systems, the CIA should include sampling of the current or reasonably anticipated prompts, as the prompts can vary and can inject or spur discrimination. *See, e.g.*, Alex Tamkin et al., Evaluating and Mitigating Discrimination in Language Model Decisions 12 (Dec. 6, 2023) (unpublished manuscript), https://perma.cc/LM5S-2YRZ.

49. C*f.* Press Release, U.S. EEOC, EEOC Hearing Explores Potential Benefits and Harms of Artificial Intelligence and Other Automated Systems in Employment Decisions

footnote continued on next page

for housing algorithms,[50] and federal or state courts for risk-assessment instruments.[51] A CIA should be required for any algorithmic system used in decisions materially affecting individuals' liberty or access to employment, housing, credit, education, immigration, criminal justice, and public benefits—the domains in which civil-rights protections have been the strongest.[52] Current AI usage, such as LLM-assisted tools, can present greater challenges than previous algorithmic approaches, including the inability to discern the classifying factors and weights.

Government backing is important for compliance and transparency. The *Mobley* case (raised in the Introduction) is a cautionary tale of bias testing left to private parties. The court recently refused to provide the plaintiff with Workday's bias-testing data after accepting Workday's representations that its counsel had curated the data, the testing served as legal advice rather than a business purpose, and the data had not been disclosed to a third party or regulator.[53] The ruling creates a predictable incentive to route internal bias testing through corporate counsel, shielding it with the attorney-client privilege, and practitioner advisories already recommend doing so.[54] CIAs, with the force of law, would impose obligations independent of litigation. Otherwise, developers can measure how their systems perform across protected classes without disclosing much, and plaintiffs must prove proxy discrimination without this significantly probative evidence, i.e., the developer's own measurement of the very entanglement this Essay proposes to scrutinize and regulate.

If federal agencies (and their state analogs) are unwilling or unable to conduct or require pre-deployment CIAs, an alternative would be to encourage private parties to conduct their own CIAs, perhaps in conjunction with

(Jan. 31, 2023), https://perma.cc/5BSK-5BFT (soliciting testimony to discern ways to prevent unlawful bias in employment decisions using AI or automated tools).

50. *Cf.* HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 91 Fed. Reg. 1475 (Jan. 14, 2026) (to be codified at 24 C.F.R. pt. 100) (recounting HUD regulations addressing disparate impact but then proposing to eliminate those regulations and defer to the courts).

51. *Cf.* Garrett & Monahan, *supra* note 20, at 448-54 (discussing court use of risk-assessment instruments).

52. I do not mean this listing to be completely exhaustive.

53. Mobley v. Workday, Inc., No. 23-cv-00770, 2026 WL 1510537, at *4 (N.D. Cal. May 29, 2026). The court noted, however, that the underlying data might still be discoverable.

54. *See* Gerald L. Maatman, Jr., Adam D. Brown & Elizabeth G. Underwood, *California Federal Court Clarifies Limits on AI Bias Testing and Applicant Data Disclosure in Mobley v. Workday*, DUANE MORRIS CLASS ACTION DEF. BLOG (June 2, 2026), https://perma.cc/WE52-C7EJ (advising that companies using AI in hiring "should structure their bias-testing under the direction of legal counsel to preserve attorney-client privilege"); Jesika Silva Blanco, Susana Medeiros & Susan Linda Ross, *Behind the Privilege Shield: Safeguarding AI Bias-Testing Data in Employment Decisions*, NORTON ROSE FULBRIGHT INSIDE TECH L. (June 24, 2026), https://perma.cc/22N5-BV3H (advising employers to ensure that counsel directs the selection and curation of data used in bias testing).

independent auditors or consultants to reduce bias and enhance expertise. A possible incentive could be that a sufficient CIA would serve as a defense or evidence in favor of the private party against later litigation over the classification's impact.[55] Similarly, Congress or the states could authorize financial incentives and research funding for fair AI development and usage.

As to the CIA process, the assessment process should generally include three phases. The first is a disclosure phase, in which the deploying entity (e.g., big business, probation department) provides the assessing agency with the system's training data, feature set, model architecture, and validation results. The second is an analysis phase, in which the agency (or an independent auditor) evaluates the system's classifications against the mutual information test (or a similar test) and conducts a prospective analysis of feedback and other effects. The third phase is a public participation phase, in which affected communities, civil rights organizations, and technical experts can comment on the assessment and raise concerns.[56]

The European Union's (EU) AI Act arguably provides a useful comparative model. The Act classifies AI systems by risk level and imposes conformity assessments on high-risk systems before they can be placed on the market.[57] Although the EU framework of course is not organized around the protected-class categories of American civil rights law, its risk-based approach to pre-deployment assessment offers an institutional template that could be adapted to the American context. The CIAs proposed here add a distinct focus: specifying and grounding the assessment in the normative commitments of the Civil Rights Act and its progeny.

To be sure, many operational details remain open, and lawmakers would need to support the proposal for it to be backed by state-based sanctions, but the case for CIA consideration and development is strong. In the alternative, a CIA approach could be followed voluntarily in the absence of state action.

IV. Fairness by Design: Choice Architecture for Algorithmic Systems

The mutual information test and the classification impact assessment are both complementary mechanisms for detecting and evaluating proxy discrimination and its impact. But detection can be insufficient without a framework for correction. Therefore, this Part proposes a regime of fairness-by-design, which the behavioral economics literature on choice architecture partly informs. To be sure, many other solutions have been proposed to boost

55. I thank Justin Pidot for suggesting this potential alternative in an earlier draft.

56. S*ee, e.g.*, Ari Ezra Waldman, *Power, Process, and Automated Decision-Making*, 88 FORDHAM L. REV. 613, 628 (2019) (noting the need for experts to evaluate algorithmic decisionmaking).

57. S*ee generally* Commission Regulation 2024/1689, 2024 O.J. (L 1689).

algorithmic fairness,[58] and these deserve attention in conjunction with the contribution below.

A. Choice Architecture and Default Rules

As Thaler and Sunstein's well-known work on choice architecture has indicated, the structure of the decision environment significantly influences the choices individuals make.[59] A default rule (e.g., the option if the decisionmaker takes no affirmative action) is a powerful tool of choice architecture, because many or even most decisionmakers accept defaults rather than actively choosing alternatives.[60] This simple insight has been applied in several policy contexts, from retirement savings (default enrollment in 401(k) plans) to organ donations (default consent).[61]

This Essay extends this insight to the design of algorithmic systems. Currently, the default design choice for most machine learning models is to optimize for predictive accuracy without fairness constraints.[62] This default produces models that exploit whatever statistical relationships exist in the data, including relationships reflecting historical discrimination. The result, perhaps predictably, is algorithmic systems with discriminatory outputs. The design environment, in other words, essentially nudges developers toward discriminatory systems by making fairness an afterthought, which must be affirmatively chosen later rather than something built into the default.[63]

B. Improving the Default

This Essay proposes that this default be inverted. For an algorithmic decisionmaking system used in a civil rights domain, such as employment, housing, credit, education, criminal justice, and public benefits, the default

58. S*ee, e.g.*, Deborah Hellman, *Algorithmic Fairness*, *in* STAN. ENCYCLOPEDIA PHIL. (July 30, 2025), https://perma.cc/H8QB-9262 (surveying various fairness proposals); Marcello Di Bello & Ruobin Gong, *Informational Richness and Its Impact on Algorithmic Fairness*, 182 PHIL. STUD. 25, 49 (2025) (concluding that informational richness would improve performance of predictive algorithms); Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 818 (2020) (arguing that the use of protected characteristics within algorithms can improve accuracy and fairness); Xiaomeng Wang, Yishi Zhang & Ruilin Zhu, *A Brief Review on Algorithmic Fairness*, 1 MGMT. SYS. ENG'G 1, 3-7 (2022) (reviewing algorithmic fairness definitions).

59. S*ee* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 1-14 (2008).

60. S*ee* Cass R. Sunstein, *The Ethics of Nudging*, 32 YALE J. ON REGUL. 413, 417-20 (2015).

61. S*ee id.* at 419.

62. S*ee, e.g.*, MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN 69-84 (2020) (noting the push for accuracy and fairness-accuracy tradeoffs).

63. C*f.* Prince & Schwarcz, *supra* note 11, at 1306-07 (proposing to "flip the default" of permitting any variable to permitting only pre-approved variables).

design specification should include a fairness consideration or constraint. Specifically, the system should satisfy a mutual information ceiling for all protected characteristics. Failing this, the deploying entity should have to demonstrate, in part through the classification impact assessment process above, a compelling interest in a higher level of informational entanglement and that no less discriminatory alternative design achieves the objective.

This default-fairness approach has several advantages over the current framework. First, it shifts the burden of justification. Under current law, the plaintiff bears the burden of establishing disparate impact and overcoming the business necessity defense.[64] Under a default-fairness regime, the deploying entity bears the burden of justifying any departure from the fairness default. This reversal is appropriate because the deploying entity has access to the model's design specifications, training data, and performance metrics—information that plaintiffs typically cannot access due to proprietary claims and limited resources.[65]

Second, the default-fairness approach applies the behavioral insight that defaults are sticky.[66] If fairness is the default, many developers will simply accept it rather than affirmatively seeking to depart from it. Moreover, the default signals what is normatively expected and reduces the cognitive and organizational costs of compliance. The default creates a gravitational pull toward fairness, just as default enrollment creates a gravitational pull toward retirement savings.

Third, the approach partly avoids the multiple-fairness-definitions problem that plagues the less-discriminatory-alternative prong of disparate impact doctrine. Rather than requiring plaintiffs to identify a specific alternative fairness constraint and demonstrate its superiority, the default-fairness regime specifies the constraint (here, the proposed mutual information ceiling) and requires the deploying entity to justify any departure. This simplifies the inquiry for courts and agencies while preserving the deploying entity's ability to argue for a different standard in appropriate cases. To be sure, I do not claim that mutual information is the universal metric of algorithmic fairness. The claim is narrower: when the legal concern is proxy discrimination—whether a formally neutral classification is functioning as a substitute for protected-class membership—mutual information measures the relevant informational relationship.

One downside, of course, is that this regime, which partly mirrors strict scrutiny, would need buy-in from lawmakers, courts, and relevant agencies to ensure adoption and to have teeth. For example, Congress and the states could

64. 42 U.S.C. § 2000e-2(k)(1)(A)(i).

65. S*ee, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 3-14 (2015) (discussing different degrees of secrecy).

66. THALER & SUNSTEIN, *supra* note 59, at 8-11.

require fairness design as a condition of government contracts, and courts could shift evidentiary burdens onto defendants using noncompliant AI tools. Granted, this is no easy feat. But once a state, national, or otherwise prominent example lights the way (such as the EU discussion above), that example could be adapted and replicated across the civil rights domains.

C. Constitutional Considerations

A default-fairness regime that incorporates attention to protected characteristics might, somewhat surprisingly, face constitutional challenge under the Court's recent equal protection jurisprudence. In *Students for Fair Admissions*, for example, the Court concluded that race-conscious admissions programs violate the Equal Protection Clause.[67] Would, then, a default-fairness constraint attempting to limit an algorithm's informational entanglement with race be characterized as a racial classification?

There are strong reasons to think not. The mutual information ceiling does not classify any individual by race or treat any individual differently on the basis of race. Rather, it constrains, at least potentially, the design of a classification system to limit its correlation with race.[68] This is analogous to designing a facially neutral policy with an eye toward avoiding disparate impact—an approach that the Court has not only permitted but implicitly encouraged. *Inclusive Communities* recognized that disparate impact liability encourages "removal of artificial, arbitrary, and unnecessary barriers"—a purpose that the default-fairness mandate directly serves.[69]

Moreover, the default-fairness approach operates at the level of system design, not individual treatment. No individual receives a benefit or burden because of race; rather, the system itself is designed to avoid producing outputs that are informationally redundant with race. This structural intervention is meaningfully distinguishable from the individualized racial preferences at issue in *Students for Fair Admissions*. The distinction between designing a fair system and allocating individual benefits by race is, or should be, constitutionally significant.[70] Of course, an antagonistic court might ultimately decide the issue differently, but the arguments above are formidable.

---

67. S*ee* Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll., 143 S. Ct. 2141, 2166 (2023).

68. S*ee* Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 849-52 (2020) (arguing that certain race-aware fairness constraints should withstand equal protection and statutory challenges).

69. S*ee* Tex. Dep't of Hous. & Cmty. Affs. v. Inclusive Cmtys. Project, Inc., 576 U.S. 519, 540 (2015) (quoting Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971)).

70. S*ee generally* Daniel E. Ho & Alice Xiang, *Affirmative Algorithms: The Legal Grounds for Fairness as Awareness*, 2020 U. CHI. L. REV. ONLINE *134 (suggesting constitutionally grounded approaches to mitigate algorithmic bias).

Conclusion

The protected-class framework of American civil rights law assumed a world in which discrimination operated through recognizable classification, such as an employer who refuses to hire women or a landlord who excludes Black tenants. Algorithmic systems avoid this framework by generating their own classifications—categories constructed from thousands of data features that bear no nominal relationship to protected characteristics but that reproduce the effects of prohibited discrimination (and at scale).

Addressing this challenge requires an epistemological reorientation. The law should shift its focus from the decisionmaker's use of named characteristics to the classification system's informational relationship to protected-class membership. The mutual information test proposed here (or an analogous approach) would provide a more precise, quantitative tool for making this assessment. Classification impact assessments, drawing in part on the philosophy of science's insights into the politics and consequences of classificatory systems, provide a proactive mechanism for evaluating algorithmic classifications before deployment. Finally, fairness-by-design mandates (backed by behavioral economics' insights) provide a structural framework for ensuring that algorithmic systems are designed with civil rights compliance as a default rather than an afterthought.

These proposals do not require abandoning the protected-class framework. They instead require supplementing it with adequate tools to address how classification now operates in our modern society. Our current path, in which we continue to apply a framework designed for human decision-makers to systems that classify by proxy and at scale, may effectively displace the civil rights revolution in our age of AI, and it may effectively concede public intervention to private actors (e.g., tech companies) whose tools evade current regulation. As this Essay suggests, there are ways to prevent, detect, and remedy this result.