CHAPTER ONE

Resetting Antidiscrimination Law in the Age of AI

On June 21, 2022, the U.S. Attorney for the Southern District of New York brought suit against Meta Platforms[1] for discriminating against Facebook users through its housing advertisement system.[2] Meta had developed a machine learning tool that allowed advertisers to select target characteristics and display ads to users with those features.[3] Using the tool, Meta's algorithms disproportionately delivered housing ads for majority-Black zip codes to Black Facebook users, and ads for majority-white zip codes to white users.[4] In *United States v. Meta Platforms, Inc.*,[5] the government alleged that Meta had "intentionally discriminated on the basis of race" and other characteristics protected by the Fair Housing Act[6] through its design and use of these algorithms.[7]

In many ways, *Meta* foreshadowed the discrimination-related puzzles ushered in by the evolution of algorithmic decisionmaking — those enabled by artificial intelligence (AI).[8] One could imagine different remedies for the kind of algorithmic discrimination alleged in *Meta*. Meta could discontinue use of its ad algorithms entirely. It could notify users that the ads they saw were determined by algorithm-enabled targeting. It could remove race and other protected characteristics from the datasets on which the algorithms were trained, in hopes of precluding the system from "relying" on such inputs. In reality, Meta quickly agreed to deploy a new "Variance Reduction System," or VRS — a second algorithmic overlay designed to reduce certain biases of the ad tool by rebalancing its results to render it less discriminatory.[9] The solution was

---

[1] Complaint at 1, United States v. Meta Platforms, Inc., No. 22-cv-05187 (S.D.N.Y. filed June 21, 2022).

[2] *Id.* ¶ 1.

[3] *Id.* ¶ 41.

[4] *Id.* ¶¶ 84–91. Studies unrelated to the lawsuit also found that Facebook employment ads were disproportionately shown to women, while ads related to credit cards, loans, and insurance were disproportionately shown to men. *See* Sara Kingsley et al., *Auditing Digital Platforms for Discrimination in Economic Opportunity Advertising*, MD4SG, June 2020, at 8, 15–18, https://arxiv.org/pdf/2008.09656 [https://perma.cc/LZT3-Q4EP].

[5] No. 22-cv-05187 (S.D.N.Y. filed June 21, 2022).

[6] 42 U.S.C. §§ 3601–3619; Complaint, *supra* note 1, ¶ 60.

[7] Complaint, *supra* note 1, ¶¶ 60, 103–04.

[8] This chapter uses the term "AI" to refer to a class of algorithmic processes (as in, processes that use computational methods to conduct operations based on certain data and instructions) that take training data as input and use information to change and refine the underlying algorithm without explicit programming. Often, AI systems incorporate machine learning techniques, computational systems, and advanced algorithms such that AI can learn and solve problems dynamically, rather than deterministically. *See* Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 Harv. J.L. & Tech. 889, 899 (2018).

[9] Settlement Agreement at 6–8, *Meta*, No. 22-cv-05187.

to adjust the system's results until its outputs were more representative of the full population that should be eligible to receive the ad.[10]

In the two years since generative AI was launched into popular use,[11] legislators have attempted to regulate AI in all these ways and more. This Chapter surveys seventy-five proposed and enacted federal and state bills targeting AI-enabled discrimination[12] in civil contexts and identifies four dominant methods of regulation: prohibiting use, regulating process, regulating inputs, and regulating outputs. Some of these methods combat AI discrimination by treating it as a new and unique harm, but the majority shoehorn AI into existing legal frameworks.

The various approaches to AI regulation reflect a longstanding debate in technology and the law. In the mid-1990s, Judge Easterbrook argued that just as society did not need to create a new legal discipline to regulate the once-novel technology of the horse, it did not need to invent a new legal paradigm to govern "cyberspace."[13] Under this logic, AI may not merit a new legal paradigm if it is the "horse" of our day. At the same time, however, the gaps that emerge when existing antidiscrimination law is extended to AI may reflect something more profound. Responding contemporaneously to Judge Easterbrook's "law of the horse" framing, Professor Lawrence Lessig instead posited that because code changes the applicability of law in cyberspace, thinking critically about the "law of cyberspace" might teach us "something general about real space law."[14] In Lessig's frame, then, the way we seek to regulate a new form of AI-enabled discrimination might teach us something about "real antidiscrimination law" — that is, how we ought to regulate old forms of the same harm by *human decisionmakers*.

---

[10] *Id.*

[11] *See What Is Generative AI?*, MCKINSEY & CO. (Apr. 2, 2024), https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai [https://perma.cc/CE46-AZPS] (noting the launch of ChatGPT in November 2022 and the subsequent growth of generative AI).

[12] This Chapter uses the term "AI-enabled discrimination" to refer to actions that have a discriminatory effect and are effectuated, in part or in whole, using AI, algorithms, machine learning, and other similar computational processes.

[13] *See* Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL F. 207, 207. Though Judge Easterbrook popularized the phrase "law of the horse," it was originated by Dean Gerhard Casper, who stated that the University of Chicago Law School did not offer a course on the topic. *Id.* Judge Easterbrook later argued that legal principles from contract, property, and tort should be sufficient to address issues arising in cyberspace. *See id.* at 208.

[14] Lawrence Lessig, The Law of the Horse: What Cyberlaw Might Teach 10–11, 24–32, 46 (Apr. 14, 1999) (unpublished manuscript) [hereinafter The Law of the Horse (unpublished manuscript)], https://cyber.harvard.edu/works/lessig/LNC_Q_D2.PDF [https://perma.cc/2KM5-K6QB]; *see also* Lawrence Lessig, Commentary, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501, 502 (1999). Lessig also originated the idea of code as law: that code can provide the rules of interaction and enshrine its own value judgments in cyberspace and may even displace law in certain contexts as the regulator of cyberspace. *See id.* at 548. *See generally* LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999).

Of the four regulatory methods this Chapter surveys, the most common for addressing AI-enabled bias regulates an AI system's outputs.[15] This method involves requirements to test, audit, report, or adjust AI systems based on their *results*. But the corollary to this type of regulation under existing antidiscrimination doctrine is disparate impact — a regime that some fear has been rendered constitutionally suspect. Rather than concede that output-oriented AI regulations are doomed, however, we can instead use the instinct to regulate AI in this way to reflect on how our current legal regime governing both machines *and* humans is deeply incomplete.

This Chapter proceeds in three sections. Section A summarizes current antidiscrimination law and its major gaps as applied to AI-enabled decisionmaking. Section B surveys bills targeting AI that have been proposed or enacted at the state or federal level since 2022, categorizes them into four regulatory methods, and analyzes each method's corollaries to existing legal frameworks. Section C analyzes the viability of output-based regulations — the most common method for addressing AI-enabled discrimination — under the Supreme Court's decisions in *Ricci v. DeStefano*[16] and *Students for Fair Admissions, Inc. v. President & Fellows of Harvard College*[17] (*SFFA*). It then offers two paths to secure their validity, either by embracing tech exceptionalism and creating an AI carveout, or by revisiting the standard we apply to human decisionmaking. The prevalence of legislative efforts to extend disparate impact liability to AI decisionmaking favors the latter. As new regulations begin to form the "law of AI," their methods underscore how and why decisions like *Ricci* and *SFFA* should be read narrowly with regard to "real antidiscrimination law," which should allow correction of formalistically neutral AI *and* human decisionmaking systems.

A. Antidiscrimination Law Before AI

U.S. law has long had frameworks to combat intentional and systemic discrimination in the context of human decisionmaking, but those frameworks may be inadequate to address bias in the age of AI. This section outlines two pre-AI antidiscrimination regimes: constitutional protections under the Equal Protection Clause of the Fourteenth Amendment[18] (EPC) and statutory protections under Title VII of the

---

[15] Over 80% of bills that target AI-enabled discrimination regulate the outputs of AI systems. That figure rises to over 90% when bills that primarily establish commissions or working groups are excluded. For an expanded discussion, see *infra* p. 1576.

[16] 557 U.S. 557 (2009).

[17] 143 S. Ct. 2141 (2023).

[18] U.S. CONST. amend. XIV, § 1. This Chapter also uses "EPC" to refer to the equal protection guarantee applicable to the federal government through the Fifth Amendment. *See* Bolling v. Sharpe, 347 U.S. 497, 499–500 (1954).

Civil Rights Act of 1964.[19] It then notes their deficiencies when applied to AI.

The EPC guarantees "the equal protection of the laws"[20] by limiting state action that distinguishes between groups of people, particularly when done on the basis of identity characteristics like race.[21] The protective reach of the EPC, however, has been curtailed by the Supreme Court, which has interpreted the EPC as effectively barring only *intentional* discrimination.[22] This intent requirement has been criticized as difficult to determine;[23] conceptually strained when ascribed to groups like legislative bodies;[24] and counter to the purpose and "spirit" of the EPC.[25]

When it comes to AI, the intent requirement could render the EPC a dead letter. For one, *whose* intent should matter? In the human context, there is a debate over how a legislator's impermissible purpose affects the intent analysis.[26] AI magnifies this problem because the decisionmaking chain is far more complex. Is the relevant intent that of the many humans whose decisions make up the AI's training data? That of the developers and red teamers[27] who select training data and then build, test, and calibrate the AI systems? That of the users who use AI in both predictable and unforeseen ways? To further complicate matters, AI systems have disaggregated "supply chains" that may trigger different technologies and datasets each time they are activated, meaning this web of potential actors can differ with every use.[28]

What's more, even if one could say that an AI system *itself* acted with "intent," *why* did it act? The "reasoning" behind AI intent is

---

[19] 42 U.S.C. §§ 2000e to 2000e-17.

[20] U.S. CONST. amend. XIV, § 1.

[21] Kenji Yoshino, *The New Equal Protection*, 124 HARV. L. REV. 747, 756 (2011) (noting the heightened scrutiny applicable to "race, national origin, alienage, sex, and nonmarital parentage" (footnotes omitted)); *id.* at 755 & n.61.

[22] *See id.* at 763–64 (citing Washington v. Davis, 426 U.S. 229, 245–47 (1976) (holding that heightened scrutiny does not apply in the absence of "discriminatory purpose," *Davis*, 426 U.S. at 247)).

[23] *Davis*, 426 U.S. at 253–54 (Stevens, J., concurring).

[24] *Id.* at 253; *see* John Hart Ely, *Legislative and Administrative Motivation in Constitutional Law*, 79 YALE L.J. 1205, 1219–20 (1970) (noting that courts and litigants would be unlikely to ascribe intent to a legislature unless that intent is "shared by a majority of the decision-makers," *id.* at 1220).

[25] Khiara M. Bridges, *The Supreme Court, 2021 Term — Foreword: Race in the Roberts Court*, 136 HARV. L. REV. 23, 104, 103–04 (2022).

[26] *See* Ely, *supra* note 24, at 1219–20 (noting that litigants are likely to bring claims only if a majority of legislators share an impermissible intent). *See also generally* W. Kerrel Murray, *Discriminatory Taint*, 135 HARV. L. REV. 1190 (2022) (examining how past impermissible purposes might "taint" the intent of subsequent legislative bodies).

[27] *See* Exec. Order No. 14,110, 3 C.F.R. 657, 660 (2024) (defining "AI red-teaming" as a "structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI").

[28] Jennifer Cobbe et al., *Understanding Accountability in Algorithmic Supply Chains*, *in* ASS'N FOR COMPUTING MACH., PROCEEDINGS OF THE 6TH ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1185, 1188, 1190 (2023).

becoming increasingly impossible to discern,[29] creating what some scholars have termed the "Black Box Problem."[30] Professor Yavar Bathaee and others have argued that the black box breaks down the EPC's intent test entirely;[31] Bathaee argues that AI should instead be governed by alternative tests like strict liability or negligence, depending on the risk context.[32] Another option is to *infer* intent from the effects of a policy or system, but the Supreme Court has increasingly eschewed such inferences in the absence of affirmative intent.[33] One could instead apply principles like respondeat superior liability when AI acts on a human's behalf;[34] at least one federal court has held that an AI system vendor can be liable for a hiring decision under such an agency liability theory.[35] Yet another proposed solution is to impose ex ante design and transparency requirements.[36]

The second antidiscrimination regime is Title VII, a federal statutory scheme that prohibits employment discrimination on the basis of "race, color, religion, sex, or national origin."[37] Though Title VII is limited to employment,[38] other regimes provide similar antidiscrimination

---

[29] *See, e.g.*, Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1089–90, 1094 (2018); Zachary C. Lipton, *The Mythos of Model Interpretability: In Machine Learning, The Concept of Interpretability Is both Important and Slippery*, ACM QUEUE, May–June 2018, at 12–15. Even where AI companies have insight into the data on which a system has trained (and thus potentially relied), companies may resist sharing such information, which is treated as akin to a "trade secret." *See* Martin Coulter, Reuters, *The EU's AI Act Raises Questions About Data Transparency and Trade Secrets*, FAST CO. (June 13, 2024), https://www.fastcompany.com/91140527/ai-act-european-union-data-transparency-training-models-trade-secrets [https://perma.cc/6XXN-RUX2].

[30] Bathaee, *supra* note 8, at 894.

[31] *Id.* at 895.

[32] *Id.* at 931–36. Bathaee argues for a framework that scales liability not only to risk but also to a system's transparency and degree of supervision. *Id.*

[33] *See* Yoshino, *supra* note 21, at 764 (discussing Personnel Administrator v. Feeney, 442 U.S. 256 (1979), which held that discriminatory intent is not present unless a decision-maker follows "a particular course of action at least in part 'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group," *id.* at 279).

[34] *See* Ian Ayres & Jack M. Balkin, *The Law of AI Is the Law of Risky Agents Without Intentions*, U. CHI. L. REV. ONLINE (Nov. 27, 2024), https://lawreview.uchicago.edu/sites/default/files/2024-11/Ayres_Balkin_Law%20of%20Risky%20Agents.pdf [https://perma.cc/N7QU-HPUT]; *see also* Jonathan L. Zittrain, *We Need to Control AI Agents Now*, THE ATLANTIC (July 2, 2024), https://www.theatlantic.com/technology/archive/2024/07/ai-agents-safety-risks/678864 [https://perma.cc/SU5B-B4XD] (discussing methods for regulating "AI agents").

[35] *See* Mobley v. Workday, Inc., No. 23-cv-00770, 2024 WL 3409146, at *5–7 (N.D. Cal. July 12, 2024). Workday provides algorithmic hiring software to employers. *Id.* at *7. In this partial grant and partial denial of Workday's motion to dismiss, Judge Lin held that Workday could be subject to liability because the algorithm it provided acted as an "agent" of its clients, thus making Workday "an 'employer' for purposes of Title VII, the ADEA, and the ADA." *Id.* at *6–7, *11.

[36] Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1875, 1943–47 (2020). *But see* Bathaee, *supra* note 8, at 929 (rejecting transparency requirements as innovation stifling).

[37] 42 U.S.C. § 2000e-2(a)(1)–(2).

[38] *Id.* § 2000e-2(a).

protections in housing,[39] credit,[40] and federal funding.[41] Unlike the EPC, Title VII is not limited to intentional discrimination: Instead, it creates a burden-shifting framework under which employees can challenge practices that have disparate impacts on protected classes.[42] If, for example, an employee makes a prima facie showing that a practice leads to disproportionate hiring[43] of men over women, then an employer must show that the practice is "job related" and "consistent with business necessity."[44] Even if the employer meets that showing, it can still be held liable if it refused to adopt a less discriminatory "alternative" that would satisfy its business needs.[45]

Disparate impact liability has an obvious appeal in the AI context. While an AI system's intent may be difficult or impossible to discern, its outcomes are readily observable.[46] Yet scholars have expressed concern that disparate impact under Title VII may be ineffective to combat AI-enabled discrimination. First, the burden-shifting framework may not effectively restrict "proxy discrimination" because employers can show that an AI system is designed to optimize for factors rooted in "business necessity."[47] And as the use of AI becomes ubiquitous,[48] it may be difficult to argue that a non-AI system could also adequately meet an employer's legitimate business needs.[49]

Second, disparate impact doctrines may be on uncertain constitutional footing. In 2009, the Supreme Court laid down a rather fine line between *adherence to* and *violation of* statutory antidiscrimination regimes in *Ricci*. At issue was a City's decision to throw out the results

---

[39] Fair Housing Act, 42 U.S.C. §§ 3601–3619 (prohibiting housing discrimination on the basis of "race, color, religion, sex, familial status, or national origin," *id.* § 3604(a)).

[40] Fair Credit Reporting Act, 15 U.S.C. §§ 1681–1681x; *id.* § 1691(a)(1) ("race, color, religion, national origin, sex or marital status, or age").

[41] Title VI of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000d to 2000d-7 ("race, color, or national origin," *id.* § 2000d).

[42] *Id.* § 2000e-2(k).

[43] *See 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*, EQUAL EMP. OPPORTUNITY COMM'N (May 18, 2023), https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial [https://perma.cc/9N82-FJGL] [hereinafter EEOC Guidance] (describing the "four-fifths rule," a "general rule of thumb" for showing "evidence of discrimination" under Title VII's disparate impact framework).

[44] 42 U.S.C. § 2000e-2(k)(1)(A)(i).

[45] *Id.* § 2000e-2(k)(1)(A)(ii); *see* Ricci v. DeStefano, 557 U.S. 557, 589–90 (2009).

[46] *See supra* notes 26–32 and accompanying text.

[47] Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1305 (2020) (arguing that companies that use AI will "typically have little problem showing that [use of AI] is consistent with business necessity . . . because, by definition, proxy discrimination helps the AI predict a legitimate objective: the target variable it is programmed to optimize").

[48] *See* Pavithra Mohan, *70% of Companies Will Use AI for Hiring in 2025, Says New Study*, FAST CO. (Oct. 31, 2024), https://www.fastcompany.com/91220282/70-of-companies-will-use-ai-for-hiring-in-2025-says-new-study [https://perma.cc/3NEE-U4YU].

[49] *See* Prince & Schwarcz, *supra* note 47, at 1305 (describing AI as "uniquely effective at optimizing [its] programmed objective").

of its fire department's promotion eligibility exam after the City observed significant racial disparities in the results, out of fear it could be held liable for disparate racial impact.[50] A group of mostly white firefighters then sued the City under Title VII and the EPC, arguing that the decision to throw out the results amounted to intentional discrimination.[51] The Supreme Court ruled for the plaintiffs,[52] leaving unclear whether disparate impact regimes might be invalid as a form of unconstitutional intentional discrimination[53] — doubts that have only grown stronger following *SFFA*.[54]

In summary, the core machinery of constitutional and statutory antidiscrimination law begins to fray when applied to AI, either because AI has no discernable intent, or because AI-enabled decisionmaking may be defensible under disparate impact regimes. The next section outlines how proposed and enacted AI regulations have attempted to address these shortcomings — at times by bringing AI within the reach of existing regimes, and at others by creating novel methods for combatting AI-enabled discrimination.

B. A Typology of "AI Regulations"

The recent explosion of AI into our national consciousness has been accompanied by a flurry of legislative action. Legislatures have enacted dozens of AI laws;[55] the Biden Administration issued several executive orders;[56] and several federal agencies have announced their intent to

---

[50] *Ricci*, 557 U.S. at 562–63. White candidates' exam results rendered them eligible for promotion at roughly twice the rate of their Black and Hispanic counterparts. *See id.* at 586.

[51] *Id.* at 562–63.

[52] *Id.* at 593; *see also id.* at 585 (holding that an employer in the City's position may avoid Title VII disparate treatment liability only if there is a "strong basis in evidence" that disparate impact liability would have arisen absent intervention).

[53] *See* Martha Minow, *Equality, Equity, and Algorithms: Learning from Justice Rosalie Abella*, 73 U. TORONTO L.J. 163, 166–67 (2023); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 694 (2017). The impact of *Ricci* on AI regulations is discussed in section C.1, *infra* p. 1579.

[54] *See, e.g.*, *Standing Up for the Rule of Law: Ending Illegal Racial Discrimination and Protecting Men and Women in U.S. Employment Practices: Hearing Before the H. Comm. on Oversight and Accountability*, 118th Cong. 7–8 (2024) (statement of Jonathan Berry, Managing Partner, Boyden Gray PLLC).

[55] The vast majority of AI-related enactments have been at the state level. *Compare* Bill Kramer, *AI Legislation, By the Numbers*, MULTISTATE.AI (May 17, 2024), https://www.multistate.ai/updates/vol-27 [https://perma.cc/D4M9-TBGF] (state), *with* IAPP RSCH. & INSIGHTS, GLOBAL AI LAW AND POLICY TRACKER 26–27 (2024), https://iapp.org/media/pdf/resource_center/global_ai_legislation_tracker.pdf [https://perma.cc/LLL5-7GRJ] (federal).

[56] IAPP RSCH. & INSIGHTS, *supra* note 55, at 26 (listing executive orders); Exec. Order No. 14,141, 90 Fed. Reg. 5469 (Jan. 17, 2025). The Biden Administration appeared to emphasize process- and effects-based models for mitigating AI bias. *See* Exec. Order No. 14,110, 3 C.F.R. 657, 683 (2024) (directing the Secretary of Health and Human Services to promote notice, human review, and auditing). By contrast, the European Union has emphasized model development and predeployment testing. *See High-Level Summary of the AI Act*, EU ARTIFICIAL INTELLIGENCE ACT (May 30, 2024), https://artificialintelligenceact.eu/high-level-summary [https://perma.cc/P6L6-6EB3].

enforce existing law against AI-assisted violations.[57] This Chapter focuses on a subset of this activity — proposed and enacted state and federal legislation — to examine the intuitive response not just to AI, but also to discrimination more broadly.

This Chapter draws from a survey of seventy-five bills — sixty-two state and thirteen federal — introduced between January 2022 and September 2024.[58] Of these, forty-four were expressly directed toward mitigating AI-enabled bias.[59] The survey focuses on civil regulations of AI in consumer-facing sectors in which antidiscrimination principles are relevant, including healthcare, education, employment, housing, technology, financial services, and consumer protection. It includes legislation that primarily aims to establish AI commissions or working groups, as well as some regulations related to internet and data privacy that also cover AI systems. It excludes bills that encourage the adoption of AI by governmental actors[60] and those that target use of AI in criminal contexts.[61]

This Chapter identifies four primary methods by which surveyed bills and laws target the development and use of AI: (a) prohibiting certain use cases; (b) regulating the process by which AI is used; (c) regulating the inputs of AI systems; and (d) regulating the outputs of AI systems. Though these four methods are analyzed as conceptually distinct, they are not mutually exclusive in application — surveyed bills frequently use more than one method to target AI.

---

[57] *See* CFPB ET AL., JOINT STATEMENT ON ENFORCEMENT EFFORTS AGAINST DISCRIMINATION AND BIAS IN AUTOMATED SYSTEMS 1–3 (2023), https://files.consumerfinance.gov/f/documents/cfpb_joint-statement-enforcement-against-discrimination-bias-automated-systems_2023-04.pdf [https://perma.cc/KX6Q-BPRB].

[58] Data are drawn from two primary sources. First, we reviewed results from Westlaw searches of state and federal proposed and enacted legislation containing keywords "AI," "artificial intelligence," and/or "machine learning" between January 2022 and August 2024. Second, we cross-referenced those results against a tracker of AI-related legislation compiled by the National Conference on State Legislatures. *See Artificial Intelligence 2024 Legislation*, NAT'L CONF. OF STATE LEGISLATURES (Sept. 9, 2024), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation [https://perma.cc/U6P7-KFCD]. For purposes of analyzing overall trends in methods of regulation, bills that had failed as of December 2024 were included.

[59] This subset of bills is comprised of legislation that contains, in its text, language that discusses AI discrimination and bias.

[60] *See, e.g.*, Assemb. 477, 220th Leg., Reg. Sess. (N.J. 2022).

[61] This includes bills regulating AI in criminal procedure, *see, e.g.*, ALA. CODE § 15-10-111 (2022) (prohibiting identification by facial recognition technology, which could be enabled by AI, as the sole basis of probable cause for arrest), and those incorporating AI into substantive criminal law offenses, *see, e.g.*, H.B. 1373, 2023 Gen. Assemb., Reg. Sess. (Pa. 2023) (criminalizing the dissemination of AI-generated impersonations without consent).

Table 1: Methods of Surveyed Legislation

Regulatory Method	Description	Legal Analogue	Possible Weaknesses
Prohibition Regulation	Total or conditional prohibitions on certain uses of AI	Strict liability	Stifles innovation
Process Regulation	Procedural requirements, including notice, appeals, and human review or decisionmaking	Due process; constructive knowledge	Provides only individual remedies; fails to establish intent
Input-Based Regulation	Prohibitions on (or selective permissibility of) using certain classes of data and their proxies in AI decisionmaking	Disparate treatment	May insulate proxy discrimination
Output-Based Regulation	Testing or auditing of AI systems for differential impact, either before or after the system is deployed	Disparate impact	Provides unclear guidance on implementation; may face constitutional challenges

1. Prohibition Regulation. — Prohibition is the most categorical method of regulation and simply bars the use of AI in certain contexts. Many prohibition regulations are targeted at contexts where biased decisionmaking may be of concern. Proposed legislation in Colorado, for example, attempted to prohibit landlords from using any algorithmic system to determine residential rental rates.[62] A proposed Rhode Island bill would have prohibited the use of facial recognition and AI-enabled algorithms in decisionmaking related to gambling.[63] And legislation in South Carolina would have prohibited web application operators from using automated decision systems (ADSs) to curate content for users under eighteen.[64]

Prohibition regulations reflect an intuition that certain decisions are so significant that AI systems cannot be used in the process under any circumstances. Bills often refer to these as "consequential decisions" that materially affect the provision, denial, or cost of a critical service

---

[62] H.B. 24-1057, 74th Gen. Assemb., 2d Reg. Sess. (Colo. 2024). The bill was cast as an unfair trade practice regulation and did not mention housing discrimination. *See id.*

[63] S. 0146, 2023 Leg., Jan. Sess. (R.I. 2023).

[64] S. 404, 2023–2024 Gen. Assemb., 125th Sess. (S.C. 2023).

or opportunity, such as housing, employment, or education.[65] Thus, even as AI tools are used to assist decisionmaking in an increasing number of contexts, perhaps we fear overreliance on technology[66] or believe that some choices must be driven by human reasoning. Professor Aziz Huq and others have conceptualized this as "a right to a human decision" when individuals are "assigned a benefit or a coercive intervention."[67] Prohibitions on AI use codify a right to a human decision in certain contexts by carving decisions out from AI influence.

Prohibition regulations skirt the gaps in antidiscrimination law analyzed in section A by essentially imposing strict liability. An employer, for example, could be held liable simply for using AI for a prohibited purpose — say, for terminating employees — the employee would not have to prove intent or show that the employer's use of AI did not serve a legitimate business necessity. Prohibitions obviate the trickier aspects of AI regulation, such as identifying a responsible actor and determining their mens rea, which makes these regulations attractive from an evidentiary standpoint. That said, they are blunt instruments that stifle innovation and experimentation,[68] which might explain their relative rarity among surveyed bills. This suggests that prohibitions may be of limited utility in curbing discrimination in most contexts.

2. Process Regulation. — A finer-tuned version of prohibition is to regulate the process by which AI can be deployed. This approach has gained traction internationally[69] and within the tech sector itself.[70] Process regulations generally take one of two forms: (a) transparency requirements, such as notice, disclosure, or appeal mandates or (b) requirements of human involvement in AI-enabled decisionmaking processes.

---

[65] *See, e.g.*, Act of May 17, 2024, ch. 198, § 6-1-1701(3)(a)–(h), 2024 Colo. Legis. Serv. (West) (defining a "[c]onsequential decision" as "a decision that has a material legal or similarly significant effect on the provision or denial to any consumer of, or the cost or terms of" services including education, employment, lending, healthcare, and insurance).

[66] Technological systems can reduce human error but can also lead to "over-reliance" and "automation bias," which can create errors. Matthew Grissinger, *Understanding Human Over-Reliance on Technology*, 44 P&T 320, 320 (2019).

[67] Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 615, 615–17 (2020). Huq notes that the EU General Data Protection Regulation (GDPR) affirmatively creates such a right by enshrining "the right not to be subject to a decision based solely on automated processing, . . . which produces legal effects concerning him or her or similarly significantly affects him or her." *Id.* at 616 (quoting Council Regulation 2016/679, art. 22, 2016 O.J. (L 119) 1, 46).

[68] *See* Bathaee, *supra* note 8, at 932. This impulse is reflected in President Biden's 2023 Executive Order, which "discouraged . . . broad general bans" on agencies' use of AI. Exec. Order No. 14,110, 3 C.F.R. 657, 691 (2024).

[69] *See* COUNCIL OF EUR. COMM'R FOR HUM. RTS., UNBOXING ARTIFICIAL INTELLIGENCE 9–10 (2019), https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64 [https://perma.cc/5KUT-ZEU4].

[70] For example, Facebook's Oversight Board is a kind of "Supreme Court" that reviews appeals of certain decisions. Kate Klonick, *Inside the Making of Facebook's Supreme Court*, NEW YORKER (Feb. 12, 2021), https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court [https://perma.cc/8BRU-79WU].

(a) Transparency Requirements. — Some regulations impose additional transparency requirements when AI is used in decisionmaking. These often come in the form of mandated disclosures. For example, legislation has been proposed to require disclosure when AI is used in pricing algorithms,[71] employment decisions,[72] and healthcare.[73] Other bills would require that individuals be allowed to challenge an algorithmic determination or the data relied upon by an AI-enabled system.[74]

AI process regulations sometimes confer greater rights to individuals than they would have if the same decision were made by a human. For example, a proposed New Jersey bill would allow employees to "contest any [adverse] employment decision that results from the use of [an] automated employment decision tool" and to obtain information including, but not limited to, a written "statement of specific reasons for an adverse employment decision."[75] Under traditional at-will employment, employees have no right to know *why* adverse actions are taken against them.[76] Under this bill, use of AI would provide the employee with greater access to information than they are entitled to under the at-will regime today.[77]

Transparency requirements may be grounded in notions of due process.[78] Some have argued for creating distinct "technological due process"[79] norms in certain contexts where automated or AI-enabled systems undermine notice and opportunity to be heard.[80] Layering additional processes when AI is used in decisionmaking is thus a form of tech exceptionalism — it increases legal scrutiny when the government relies on technology that may erode due process rights.[81] And regardless of whether the government or a private actor relies on AI, process

---

71 *See, e.g.*, S.B. 1154, 2023–2024 Leg., Reg. Sess. (Cal. 2024).

72 *See, e.g.*, H. 1873, 193d Gen. Ct., Reg. Sess. §§ 3–4 (Mass. 2023).

73 *See, e.g.*, H. 1974, 193d Gen. Ct., Reg. Sess. § 3(c)–(d) (Mass. 2023).

74 *See, e.g.*, H. 1873, 193d Gen. Ct., Reg. Sess. §§ 2B, 5C (Mass. 2023) (giving employees a right to dispute the accuracy of the data relied on for the adverse action as well as the action itself).

75 Assemb. 3854, 221st Leg., Reg. Sess. § 2(f)(4) (N.J. 2024).

76 *See* Dallas Estes, *Preventing a Dystopian Work Environment: AI Regulation and Transparency in At-Will Employment*, ONLABOR (Sept. 28, 2023), https://onlabor.org/preventing-a-dystopian-work-environment-ai-regulation-and-transparency-in-at-will-employment [https://perma.cc/8GNX-DQTQ].

77 *See id.*

78 *See, e.g.*, Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1281–88 (2008). *See generally* Chris Chambers Goodman, *AI, Can You Hear Me? Promoting Procedural Due Process in Government Use of Artificial Intelligence Technologies*, 28 RICH. J.L. & TECH. 700 (2022).

79 Citron, *supra* note 78, at 1249 (italics omitted).

80 *Id.* at 1249–50, 1281–84, 1305; *see also* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 109 (2014).

81 Tech exceptionalism is often discussed in the Fourth Amendment search and seizure context. *See, e.g.*, Paul Ohm, *The Many Revolutions of* Carpenter, 32 HARV. J.L. & TECH. 357, 399, 401–03 (2019) (defining tech exceptionalism as a "deep and abiding belief in the exceptional nature of the modern technological era," *id.* at 399). Use of AI in decisionmaking in civil contexts may similarly justify increased scrutiny.

regulations may also mitigate general social distrust of AI.[82] By allowing people to challenge AI-enabled decisions or the underlying data, process rights mirror the more transparent participatory processes required in other adjudicatory settings, fostering a sense of fairness and promoting deference to legal determinations.[83]

Though transparency requirements are common,[84] they may not do much to fill gaps in antidiscrimination law. Providing notice of the use of AI, for example, does not alter one's actual intent under the EPC or an employer's business necessity claim under Title VII. And while appeals processes might provide an opportunity to identify and correct discrimination in individual cases, if such discrimination is a symptom of broader bias in an AI system, case-by-case appeals are an unsatisfying remedy.[85] At best, then, disclosures and appeals provide means of individual relief, but they seem ill-suited to rectify systemic bias in AI decisionmaking.

(b) Requirement of Human Involvement. — Other regulations require human involvement in decisions made with AI assistance. Unlike prohibitions, this method permits the use of AI but requires human engagement at certain stages, such as in the initial determination[86] or on appeal.[87] A Louisiana bill, for example, would bar healthcare providers from making decisions regarding patient care based solely on an AI determination.[88] Other healthcare bills similarly mandate that patients have the option of interacting with a human provider for medical diagnoses[89] or in patient care communications.[90]

Human involvement may bridge gaps in antidiscrimination law by imputing a kind of constructive knowledge of AI-enabled bias onto a human actor. For example, Amazon discovered in 2015 that an algorithm it had built (but not used) to screen resumes was penalizing

---

[82] *See* Cynthia Dwork & Martha Minow, *Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law*, DAEDALUS, Spring 2022, at 309, 312–13.

[83] *See id.* at 311–12 ("Trust in the fairness of legal systems increases when those affected participate with substantive, empowering choices within individual trials or panels reviewing the conduct of police and other officials." *Id.* at 312.).

[84] Roughly half of surveyed AI bills had a transparency requirement. *See, e.g.*, Act of Aug. 9, 2024, Pub. Act 103-804 § 2-102(L)(2), 2024 Ill. Legis. Serv. (West) (as introduced Feb. 17, 2023) (requiring notice); Assemb. 3854, 221st Leg., Reg. Sess. § 2(d)–(e) (N.J. 2024) (same).

[85] *See* Huq, *supra* note 36, at 1905–06 (arguing that accuracy must be judged "across the population of regulated cases" rather than through a "granular focus on . . . the isolated case," *id.* at 1906).

[86] *E.g.*, H.B. 916, 2024 Leg., Reg. Sess. § 1(A)–(B) (La. 2024); Act of July 12, 2024, ch. 209, § 5-D:4, 2024 N.H. Legis. Serv. (West).

[87] *E.g.*, S. 2419, 118th Cong. § 3(a)(1)(B)(vii)(II) (2023).

[88] H.B. 916, 2024 Leg., Reg. Sess. (La. 2024).

[89] *See, e.g.*, H.B. 1002, 103d Gen. Assemb., Reg. Sess. (Ill. 2023).

[90] *See, e.g.*, Act of Sept. 28, 2024, ch. 848, 2024 Cal. Legis. Serv. (West) (creating disclosure requirements when AI is used to generate healthcare communications, but relaxing those requirements when communications are reviewed by a human provider).

resumes from female applicants.[91] If Amazon had put that algorithm to use, the company could hypothetically have defended against a disparate impact claim by arguing that its custom-built algorithm was rooted in business necessity and that the company lacked an effective alternative. That might absolve Amazon of liability, but it would surely be a less convincing argument if someone at Amazon *knew* that its algorithm was biased, given that an obvious "alternative" would be to adjust the algorithm in light of that knowledge.[92] The same logic extends to Meta's ad deployers after they became aware of the biased operation of Meta's housing algorithm.[93] Adding a human into the loop[94] — that is, a person who is aware of an algorithm's decisions over time — could thus make the deployer constructively aware that the AI's "target variable" is biased.[95] This knowledge could prompt investigation into alternatives.[96] Requiring human involvement could also provide a basis for other types of liability, such as under an agency theory, as knowledge of a system's bias increases its foreseeable risks.[97]

Notably, however, human in the loop[98] solutions do not change the outcome of a constitutional discriminatory intent analysis. The Supreme Court has squarely rejected the idea that disparate impact claims are actionable under the EPC, even if there are known disparate effects.[99] In the Amazon example, then, even if an employee began to notice that the tool selected more resumes from men, that would not be enough to transform inadvertent bias into intentional discrimination.[100]

3. Input-Based Regulation. — A third category of regulation targets the inputs of AI systems, typically by barring use of certain classes of

---

91 Jeffrey Dastin, *Insight — Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018, 8:50 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G [https://perma.cc/8NB4-2N8L].

92 *Cf. supra* note 45 and accompanying text.

93 *See supra* notes 9–10 and accompanying text.

94 *See* Rowena Rodrigues, *Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities*, J. RESPONSIBLE TECH., Dec. 2020, https://www.sciencedirect.com/science/article/pii/S2666659620300056 [https://perma.cc/54ZV-RQYB].

95 *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 706 (2016) ("If the target variable is not sufficiently job related, a business necessity defense would fail, regardless of the fact that the decision was made by algorithm.").

96 *See* EEOC Guidance, *supra* note 43 (providing guidance on application of disparate impact framework to algorithms and noting that an employer should consider whether "another test . . . would be comparably as effective").

97 *See* Ayres & Balkin, *supra* note 34, at 2 (arguing that humans could be held liable for "failing to exercise due care in supervising or training" AI agents and for "hosting an AI program . . . that creates foreseeable risks of harm").

98 *See* Rodrigues, *supra* note 94.

99 *See* Pers. Adm'r v. Feeney, 442 U.S. 256, 279 (1979).

100 *See* Barocas & Selbst, *supra* note 95, at 700 ("[T]o be found liable under current [disparate treatment] doctrine, the employer would likely both have to know that this is the specific failure mechanism of the model *and choose it based on this fact*." (emphasis added)).

data.[101] An Illinois law, for example, would have barred the use of AI systems to consider an applicant's race to reject an applicant for employment, assign risk factors to an applicant's creditworthiness, or take adverse action on a credit decision.[102] A California law requires healthcare-related AI systems to rely only on patients' medical histories and clinical information, presumably barring models trained on demographic or other nonclinical information.[103]

Input-based regulations attempt to create characteristic-blind AI systems to prevent discrimination. The prohibited inputs are typically protected characteristics (most often, race) or their proxies (such as income, education, or zip code).[104] There is a robust academic debate over whether requirements to create such mechanistically "race-neutral" algorithms mitigate, exacerbate, or themselves amount to discrimination under the EPC.[105] Some have argued that input-based practices may be necessary to avoid "serious constitutional concerns."[106] Because reliance on a protected characteristic may be problematic even when it is just one factor among many,[107] one might conclude that protected characteristics must be entirely withheld from AI systems,[108] particularly if a system is a black box.[109] Others have argued, however, that input restrictions are highly unlikely to be effective at combatting AI-enabled discrimination in practice. Professors Crystal Yang and Will Dobbie

---

[101] Input-based regulations were relatively infrequent among recent legislation — only two of the seventy-five surveyed bills contained input restrictions. But such methods have historically been used to target bias in healthcare, *see* Prince & Schwarcz, *supra* note 47, at 1264 (discussing the Patient Protection and Affordable Care Act and the Genetic Information Non-discrimination Act), as well as AI-enabled bias in criminal law, *see infra* note 104.

[102] Act of Aug. 9, 2024, Pub. Act 103-804, 2024 Ill. Legis. Serv. (West) (as introduced Feb. 17, 2023). The law would have also barred using zip code "as a proxy" for protected characteristics. *Id.* § 2-102(L)(1). The version of the bill that ultimately passed, *id.* (as enacted Aug. 9, 2024), replaced this input-based restriction with an output-based one.

[103] *See* Act of Sept. 28, 2024, ch. 879, 2024 Cal. Legis. Serv. (West) (amending CAL. HEALTH & SAFETY CODE § 1367.01 (West 2022)).

[104] Though outside this survey's scope, AI-enabled criminal justice tools are frequently regulated through input restrictions and almost always exclude race and at least some proxy variables. *See* Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 330–31 (2020).

[105] *See* Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2224 (2019); Prince & Schwarcz, *supra* note 47, at 1275–76; Yang & Dobbie, *supra* note 104, at 295–97; Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1100 (2019); *cf.* Barocas & Selbst, *supra* note 95, at 719; Huq, *supra* note 36, at 1920 (noting that "[h]uman decision makers . . . are often inevitably aware of race" and arguing that, "[b]y analogy, . . . mere inclusion of race as a feature of training data should not be per se problematic" unless it "influences ultimate decisions in a constitutionally relevant way").

[106] Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 806 (2014); *see id.* 823–29; Dawinder S. Sidhu, *Moneyball Sentencing*, 56 B.C. L. REV. 671, 694–96 (2015).

[107] *See* Bostock v. Clayton County, 140 S. Ct. 1731, 1739 (2020).

[108] *See* Yang & Dobbie, *supra* note 104, at 304–07 (describing the "mainstream view," *id.* at 304, that it is likely unconstitutional to use demographic traits and socioeconomic status as factors in criminal justice sentencing algorithms).

[109] *See* Bathaee, *supra* note 8, at 901–06 (describing the "Black Box Problem," *id.* at 903).

explain that race is so strongly correlated with other factors that withholding racial data will not prevent disparate sentencing recommendations by AI systems[110] — and in fact, that withholding such data could exacerbate race-based disparities.[111]

Input-based regulations do not appear to solve the weaknesses in current antidiscrimination law, and they may in fact have the perverse effect of *insulating* AI bias from the law's reach. When it comes to discriminatory intent, those who use race-blind systems can argue that, because all protected characteristics were excluded, the AI's output simply could not have been based on those characteristics. But in fact, AI systems are far better at identifying highly correlated variables than humans are,[112] and mechanistically race-blind systems can readily find or create proxies for race even where such data has been categorically excluded from the system inputs.[113] By assuming that AI can make "characteristic-blind" decisions at all, input-based regulations thus ironically provide cover for the very kinds of invidious discrimination the law is intended to prevent.

4. Output-Based Regulation. — The final regulatory method imposes requirements based on the outputs of AI systems, typically requiring testing or auditing[114] to determine whether an AI system has biased effects.[115] Among surveyed regulations, output-based methods are the most common for targeting AI-enabled bias and are included in roughly 82% of bills with a focus on discrimination. That figure rises to 94% if bills that primarily create AI commissions or working groups are excluded.[116] Only a few regulations mandating bias audits actually define them, however;[117] a New Jersey bill would require employers to calculate an AI system's "selection rate," defined as the ratio of favorable to unfavorable employment decisions along certain protected

---

110 Yang & Dobbie, *supra* note 104, at 335–41.

111 *Id.* at 342–43; Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action*, 110 CALIF. L. REV. 1539, 1542 (2022).

112 *See* Kalev Leetaru, *A Reminder that Machine Learning Is About Correlations Not Causation*, FORBES (Jan. 15, 2019, 2:17 PM), https://www.forbes.com/sites/kalevleetaru/2019/01/15/a-reminder-that-machine-learning-is-about-correlations-not-causation [https://perma.cc/5UZB-TFMB].

113 *See* Prince & Schwarcz, *supra* note 47, at 1263–64, 1303.

114 Testing and auditing requirements are distinguished from process requirements in that the latter govern *how* an AI system may be implemented, while the former impose ex ante or ex post checks on that system.

115 *See, e.g.*, H. 114, 2023 Leg., Reg. Sess. § 1 (Vt. 2023) (requiring impact assessment); A.B. 331, 2023–2024 Leg., Reg. Sess. § 1 (Cal. 2023) (same); Artificial Intelligence Governance Act of 2024, ch. 496, 2024 Md. Legis. Serv. (West) (codified at MD. CODE ANN., STATE FIN. & PROC. § 3.5-803(e)(1) (West 2024)) (same); H.R. 5628, 118th Cong. § 3(b)(1) (2023) (same).

116 In contrast, 75% of noncommission bills with a focus on AI bias or discrimination utilize one or both types of process regulations: 67% include a transparency requirement and 32% require human involvement.

117 *Cf.* Alfred Ng, *Can Auditing Eliminate Bias from Algorithms?*, THE MARKUP (Feb. 23, 2021, 8:00 AM), https://themarkup.org/the-breakdown/2021/02/23/can-auditing-eliminate-bias-from-algorithms [https://perma.cc/8T2S-WJE5] (noting the lack of definitions and standards of what legally mandated "audits" actually entail).

characteristics, as well as its "impact ratio," or the ratio of favorable outcomes in the "protected class" compared to a "control class."[118] Both metrics rely on the sample *outputs* of AI systems to determine whether those systems are biased.

The regulations differ in what response, if any, is required once testing reveals that a system has produced disparate results. Many regulations simply require that results be disclosed — at times to the public,[119] but more often to an agency[120] or state official.[121] Some require disclosure of all impact assessment results,[122] while others require public disclosure only if biased effects are found.[123] Other regulations stipulate that systems that produce "discriminatory" results cannot be used at all.[124] Between the two extremes, some regulations prohibit AI decisions that discriminate "based on" protected characteristics[125] or "in a manner" that discriminates against a protected class.[126] In these cases, it is not always clear whether a system that produces "biased" results can *never* be used, or whether its use is prohibited only in individual instances where results are shown to be biased. Still other regulations provide little to no guidance beyond the fact that an impact assessment must be conducted.[127]

The relationship between output-based regulations and existing anti-discrimination frameworks depends on when impact assessments are conducted and what follow-up actions are required. Some regulations mandate pre-implementation *testing* before an AI system may be put

---

118 Assemb. 3855, 221st Leg., Reg. Sess. (N.J. 2024) (as adopted by Assemb. Sci., Innovation & Tech. Comm., May 16, 2024). Protected characteristics include "race, . . . national origin, ethnicity, sex, gender identity, sexual orientation, age, religion, . . . familial status, [and] disability." *Id.* at 2. The bill does not define what is meant by a "control class." If the decision made is not binary (for example, a decision to increase salary by some amount, versus a binary decision to promote or not to promote), the impact ratio is calculated as a standard deviation of the overall population outcome. *Id.* at 2–3.

119 A.B. 567, 2023–2024 Gen. Assemb., Reg. Sess. § 1 (N.Y. 2023).

120 S. 3015, 221st Leg., Reg. Sess. § 1(d)–(e) (N.J. 2024); H.R. 5628, 118th Cong. § 3(b)(1)(D) (2023).

121 B25-0114, 25th Council, Reg. Sess. § 7(b)(1) (D.C. 2023) (requirement to disclose to attorney general).

122 *Id.*

123 Artificial Intelligence Governance Act of 2024, ch. 496, 2024 Md. Legis. Serv. (West) (codified at MD. CODE ANN., STATE FIN. & PROC. § 3.5-804(b)(4) (West 2024) (requiring notification of any group determined to have been "negatively impacted" by a high-risk AI system).

124 *See, e.g.*, S. 1402, 220th Leg., Reg. Sess. §§ 2–4 (N.J. 2022).or example, a New Jersey bill would prohibit financial institutions, lenders, insurers, and healthcare providers from using "discriminatory" ADSs. *Id.* A system is "discriminatory" if it "selects individuals who are members of a protected class for participation or [services] eligibility . . . at a rate . . . disproportionate to the rate" of members outside the class. *Id.* § 2.

125 *See, e.g.*, H. 5734, 2023 Gen. Assemb., Jan. Sess. § 1(a)(2) (R.I. 2023).

126 *See, e.g.*, B25-0114, 25th Council, Reg. Sess. § 4(a)(1) (D.C. 2023).

127 *See, e.g.*, S.B. 5356, 68th Leg., Reg. Sess. § 4(2) (Wash. 2023).

into use,[128] while others require postdeployment *auditing* to evaluate the outcomes of an AI system after it has been relied upon.[129]

Pre-implementation testing can be evaluated under similar frameworks to the three other methods of regulation outlined above. Output-based regulations that prohibit the use of AI systems that fail pre-implementation testing are essentially conditional prohibitions. These regulations apply a strict liability framework, treating AI systems that fail bias testing as per se too risky for use. In contrast, regulations that require disclosure of pre-implementation testing results are akin to process and notice regulations.

Postdeployment auditing regulations, on the other hand, are analogous to both human in the loop process regulations and disparate impact regimes. A deployer may become constructively aware of a system's risks when auditing reveals that the system has produced discriminatory outputs, requiring a deployer to then take corrective action under Title VII or other disparate impact regimes.

5. Reflections. — Of the four regulatory methods outlined above, some treat AI as meriting new legal paradigms, while others treat it as a twenty-first-century "horse" that existing doctrines can readily proscribe. Prohibitions and process transparency regulations place restrictions on AI above and beyond what human decisionmakers would face when acting without AI input,[130] thereby treating AI as inflicting new harms distinct from human discrimination. By contrast, input- and output-based regulations and human involvement requirements appear to target AI discrimination as the latest manifestation of biased human decisionmaking, attempting to bring it within the reach of existing anti-discrimination schemes.[131] The next section explores whether there *should* be a presumption that AI behavior is different from human behavior and what that might imply about "real antidiscrimination law" as applied to both human and AI-made decisions.

C. The Output-Based Puzzle

Given the array of these regulations proposed in recent years, it is clear that legislators are concerned about AI-enabled discrimination. But output-based regulation, the most commonly proposed method for combatting AI bias, may face legal challenges under modern Supreme Court doctrine. This section offers two paths forward. First, treatment of AI-enabled discrimination should be carved out from any further erosion of disparate impact regimes. Second, and more conceptually, AI

---

[128] *See, e.g.*, Assemb. 3854, 221st Leg., Reg. Sess. § 2(a)(1) (N.J. 2024) (prohibiting sale or use of an ADS unless it was audited within the prior year). Pre-implementation testing is also referred to as red-teaming. *See* Exec. Order No. 14,110, 3 C.F.R. 657, 660 (2024).

[129] *See, e.g.*, Assemb. 3855, 221st Leg., Reg. Sess. § 2 (N.J. 2024); A.B. 567, 2023–2024 Gen. Assemb., Reg. Sess. (N.Y. 2023).

[130] *See supra* notes 75–77 and accompanying text.

[131] *See supra* notes 98–99, 112–13 and accompanying text.

forces the question of whether the potentially discriminatory harms it creates are truly distinct from those in a pre-AI world. If not, output-based regulations may suggest that the same antidiscrimination regime should also normatively apply to *human* action. This path suggests that disparate impact doctrine must be retained and strengthened not just for AI, but for human decisionmaking as well.

1. Disparate Impact: A Weakening Regime. — Of those bills surveyed, nearly all proposed or enacted regulations targeting AI-enabled discrimination include a testing or auditing requirement.[132] After *Ricci*, however, output-based regulations may be of uncertain constitutional validity. In *Ricci*, the City threw out its employment exam results in part to avoid disparate impact liability under Title VII.[133] The Court held that the City's decision amounted to "intentional discrimination,"[134] a characterization that raises questions as to whether disparate impact regimes that require alteration of biased outputs will be held unconstitutional.[135] Though the Court has not returned to this question since it decided *Ricci* in 2009, its 2023 decision in *SFFA* embraced an anticlassification view of the EPC that may render *any* consideration of protected characteristics legally suspect, even if one's goal is to prevent disparate outcomes.[136]

2. An AI Carveout. — If the Supreme Court invalidates or continues to narrow disparate impact regimes, one way to protect output-based regulations is to create an AI carveout. Such an approach could be justified by at least two unique challenges posed by AI. First, and most critically, output-based regulation may be mechanistically necessary in the AI context, as the lack of transparency in AI decisionmaking[137] may

---

132 Of all such surveyed bills, 82% are output-based regulations; that figure increases to 94% if bills that primarily establish commissions or working groups are excluded.

133 Ricci v. DeStefano, 557 U.S. 557, 562–63 (2009).

134 *Id.* at 585.

135 *See* Minow, *supra* note 53, at 176 ("[T]esting an algorithm on applicant data and then modifying it in light of results could expose an employer to similar jeopardy [as in *Ricci*]."); Barocas & Selbst, *supra* note 95, at 725 (noting that auditing legislation "may run afoul of *Ricci*," as "correct[ing] for detected biases" would require employers "to consider membership in the protected class," which "is inherently race-conscious").

136 *See* Jan-Laurin Müller, *Fairness in Machine Learning as "Algorithmic Positive Action*," EUR. WORKSHOP ON ALGORITHMIC FAIRNESS, June 7th to 9th, 2023, at 1, 3, https://ceur-ws.org/Vol-3442/paper-46.pdf [https://perma.cc/2WXL-L67A] (noting that "post-processing" algorithmic interventions — that is, postdeployment auditing regulations — are likely to be scrutinized most rigorously, but that "even pre- and in-processing methods" — that is, predeployment testing and real-time corrections — "will largely be considered illegal" under then-pending *SFFA*); *see also* Daniel E. Ho & Alice Xiang, *Affirmative Algorithms: The Legal Grounds for Fairness as Awareness*, U. CHI. L. REV. ONLINE (Oct. 30, 2020), https://lawreview.uchicago.edu/online-archive/affirmative-algorithms-legal-grounds-fairness-awareness [https://perma.cc/PXQ9-XQQP] (describing *Ricci* as a "sharp turn" toward an anticlassification view of the EPC that puts race-conscious corrective remedies to algorithmic bias at "serious legal risk[]").

137 *See generally* Hannah Bloch-Wehba, *Transparency's AI Problem*, KNIGHT FIRST AMEND. INST. (June 17, 2021), https://knightcolumbia.org/content/transparencys-ai-problem [https://perma.cc/7HD6-R3YA].

escape what our current adjudicative process can assess, prevent, and rectify. It is difficult, if not impossible, to detect AI-enabled bias without examining a system's outputs.[138] And while AI systems can be red-teamed or tested for bias prior to use, they may well produce unanticipated results in the "real world"[139] compared to other algorithmic and machine learning systems. Thus, examining an AI's outputs may be the only effective way to scrutinize it.[140] And other regulatory approaches have drawbacks: Flat prohibitions stifle innovation; process regulations operate at the individual level and are likely ineffective to address system-wide bias; and input-based regulations are widely considered inadequate protection against (or even a driver of) AI-enabled bias. In this world, a bar on output-based interventions could become a functional mandate to accept AI-assisted bias once a system has been deployed.

Second, a carveout could be justified on the view that AI-enabled decisionmaking is simply different from that of humans in its scale or nature. There is a sense that AI is unaccountable and that its decisions are arbitrary, which infringes on a dignitary interest of being able to attribute adverse decisions to a responsible actor and seek explanations for them.[141] And because AI systems generate predictive outcomes based on prior human decisions, there is a very real concern that AI could amplify or legitimate existing human bias[142] at a massive scale.[143] Under this view, the effects of ubiquitous and nearly costless AI decisionmaking are of a different kind or order of magnitude than New Haven's decision to toss the exam results of 118 firefighters.[144]

Both of these justifications are variations on tech exceptionalism: the idea that, when a technology is fundamentally different from what has come before, it justifies — and often forces — a change in the law.[145]

---

138 Thomas B. Nachbar, *Algorithmic Fairness, Algorithmic Discrimination*, 48 FLA. ST. U. L. REV. 509, 542–43 (2021); *see* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 708–09 (2017).

139 Kim, *supra* note 111, at 1585.

140 *See, e.g.*, Huq, *supra* note 36, at 1948 ("Ex ante regulation is necessary, but is not sufficient . . . . Designers of a machine-learning system cannot be certain before the fact of how their instrument will perform across all conceivable circumstances."); Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 981–82 (2018); Tal Zarsky, *The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 SCI. TECH. & HUM. VALUES 118, 127 (2016). *See also generally* Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503 (describing the importance of transparency and advancing a framework for understanding the role it must play in AI regulation).

141 *See* Dwork & Minow, *supra* note 82, at 312.

142 Minow, *supra* note 53, at 163–64.

143 *See* Leonardo Nicoletti & Dina Bass, *Humans Are Biased. Generative AI Is Even Worse*, BLOOMBERG (June 9, 2023), https://www.bloomberg.com/graphics/2023-generative-ai-bias [https://perma.cc/572M-LNGB].

144 *See* Ricci v. DeStefano, 557 U.S. 557, 562 (2009).

145 Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 553–58 (2015).

However, creating an AI exception from the disparate impact regime may create anomalous results. A carveout would bifurcate decisions made by AI versus those made by humans and apply different antidiscrimination standards and remedies to each. Putting aside whether we can practically draw the line between human and AI decisions, a tech exceptionalist approach could result in a legal regime that is much more protective of antidiscrimination principles when AI is involved than when it is not. Even if this outcome may be normatively desirable, it ultimately elides the question of *why* we believe disparate decisionmaking should be corrected in the first place. The following section provides an alternative doctrinal and conceptual treatment of AI-enabled bias that attempts to resolve this more fundamental question.

3. An AI Reflection. — The justifications for an AI carveout presuppose that there *are* fundamental differences between AI and human decisionmaking. Otherwise, there would be no need for additional protections when AI is involved.[146] But is it really true that AI systems are more harmful or pernicious when it comes to biased decisionmaking than humans are? Advocates sometimes point to AI's ability to *reduce* bias[147] because humans make irrational and inconsistent decisions based on biased priors.[148] And mechanistically, if most AI decisions simply reflect the human decisions on which they are trained, they act as highly accurate mirrors of human bias rather than independent sources of discrimination.[149]

With this in mind, there is an alternative way to treat AI-enabled discrimination that does not require a carveout. The emergence of AI was a blank slate onto which legislators could design a normatively desirable antidiscrimination regime. Thus, our instinctual response to regulating AI reflects how the law ought to mitigate against discriminatory harms effectuated by *any* actor, whether machine or human. In other words, if AI is a mirror reflecting biased human judgments, then AI regulations are also a refraction of our desired legal remedies for *human* discrimination.

---

146 The regulations in fact presuppose that AI is *more* discriminatory than humans. If not, we would see the heightened regimes applied to AI — including strict liability, requirements of race-blindness, and additional procedural safeguards — applied to human decisionmakers under existing law.

147 *See, e.g.*, Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237, 240 (2018) (noting some beneficial results when machine learning is used to assist judges in making bail decisions). It has also been argued that AI could make *human* bias easier to detect, and that AI could thus serve as a "potential positive force for equity." Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 113, 113–14 (2018); *see also* Cass R. Sunstein, *Governing by Algorithm? No Noise and (Potentially) Less Bias*, 71 DUKE L.J. 1175, 1187–90 (2022).

148 *See* Sunstein, *supra* note 147, at 1203–04.

149 *See* Arvind Sanjeev, *Bias in AI Is a Mirror of Our Culture*, MEDIUM: UX COLLECTIVE (Mar. 16, 2023), https://uxdesign.cc/bias-in-ai-is-a-mirror-of-our-culture-3607bd795c57 [https://perma.cc/35GT-ZCAH]. *See generally* Mayson, *supra* note 105 (arguing that machine predictions "project the inequalities of the past into the future," *id.* at 2218).

The prevalence of output-based regulations suggests that we find *insufficient* a regime that cannot correct for disparate outcomes of formalistically neutral systems. For one, the frequency of output-based regulations — as compared with process regulations, which provide largely individual remedies — indicates an appetite for systemic correction in addition to fair outcomes in individual cases. Moreover, regulations seem to target biased outcomes as an important heuristic for biased decisionmaking. We appear to expect AI regulations to protect against disparate outcomes, regardless of whether they result from a specific intent to discriminate. Conversely, input-based regulations that functionally codify an anticlassification approach are relatively uncommon.

4. A Path Forward. — If AI regulations reflect what we instinctually want antidiscrimination law to accomplish, then acquiescing to a dampened disparate impact regime while carving out an "exception" for AI is a move in the wrong direction. Instead of asking whether AI regulations can address harms unique to AI, the question should be whether the law can address harms that are the *same* regardless of whether they are created by AI or by human decisions. While AI may operate at a larger scale than humans, the core issue that output-based regulations target is the same as that which disparate impact regimes have long sought to address: namely that, absent intervention, facially neutral tools may perpetuate and entrench systemic biases.[150] Given this, there are two ways to read *Ricci* narrowly to allow not only the continued operation of output-based AI regulations, but also their human analogues in disparate impact regimes.

One way to limit *Ricci* is in the timing of intervention. In *Ricci*, the Court expressly stated that an employer may, "*before administering* a test or practice," consider "how to design that test or practice . . . to provide a fair opportunity for all individuals, regardless of their race."[151] Several scholars have thus read *Ricci* to permit race-aware practices in the design and testing phases,[152] which would render output-based regulations with predeployment testing requirements safe from scrutiny. Another option is to read *Ricci* not as barring *all* postdeployment interventions, but instead as prohibiting only the "disrupt[ion of] legitimate, settled expectations" of *actual applicants*.[153] After all, the City in *Ricci*

---

[150] *See* Griggs v. Duke Power Co., 401 U.S. 424, 430 (1971) ("Under [Title VII], practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to 'freeze' the status quo of prior discriminatory employment practices.").

[151] Ricci v. DeStefano, 557 U.S. 557, 585 (2009) (emphasis added).

[152] *See, e.g.*, Kim, *supra* note 111, at 1563; Kroll et al., *supra* note 53, at 695; *cf.* Minow, *supra* note 53, at 176 (describing the City's "mistake" as "administering a test to actual applicants and then discarding the results"). Professor Martha Minow additionally challenges the presumption that a race-blind approach would in fact be neutral, noting that "there is no neutral starting point," as choices in the design of an AI system all "reflect judgments." *Id.*

[153] *See* Kim, *supra* note 111, at 1559, 1559–63 (making this argument). *But see* Barocas & Selbst, *supra* note 95, at 725–28 (arguing that alteration of an AI-enabled model after an audit reveals biased results could "run afoul of *Ricci*," *id.* at 725).

did not alter its exam protocols prospectively, but instead threw out the results of candidates who had already taken the exam.[154] Under this view, the timing of intervention may matter only as a proxy for reliance, as there are greater reliance interests after a system has been used on actual applicants. Even postdeployment auditing regulations would be within constitutional bounds under this approach, so long as only *prospective* alterations are made to AI systems.[155]

Extending these principles to human decisionmaking is revealing. Consider, for example, affirmative action in university admissions. In *SFFA*, the Supreme Court held that race may not be used as a "plus" (or a minus) in admissions decisions.[156] Viewed through the lenses of intervention timing and reliance, *SFFA* can be understood as barring interventions that adjust for racial bias in real time by admitting or rejecting specific candidates who applied with the "legitimate expectation not to be judged on the basis of race."[157] An open question after *SFFA*, however, is whether the decision bars other admissions policies like "Top Ten Percent" plans that are designed with the intent of promoting student body diversity.[158] Under the values reflected in output-based AI regulations and a narrow reading of *Ricci*, the answer should be a resounding "no," so long as system alterations are purely prospective and do not disrupt the reliance of actual applicants. In other words, just as Meta may alter or discard an ad algorithm known to be biased against its users, so too should an admissions team be permitted to reject practices — such as emphasis on test scores above other criteria[159] — known to have disproportionate impacts on certain classes of applicants, provided that they do not affect prior applicants' legitimate expectations.[160]

---

154 *See Ricci*, 557 U.S. at 580 ("The City rejected the test results solely because the higher scoring candidates were white."); *see also* Minow, *supra* note 53, at 176.

155 Kim, *supra* note 111, at 1562; Ho & Xiang, *supra* note 136, at 152 (noting that reliance interests like those in *Ricci*, where firefighters "had spent significant time and money preparing for the exam," "are not necessarily relevant in the algorithmic context").

156 Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll., 143 S. Ct. 2141, 2165–66 (2023) (citing Grutter v. Bollinger, 539 U.S. 306, 341 (2003)).

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

158 *Compare* Fisher v. Univ. of Tex. at Austin (*Fisher II*), 579 U.S. 365, 388 (2016) (upholding a university "Top Ten Percent Plan" whose "basic purpose . . . [was] to boost minority enrollment," *id.* at 386), *with SFFA*, 143 S. Ct. at 2174–75 (emphasizing the restricted nature of the *Fisher II* Court's holding), *and* Coal. for TJ v. Fairfax Cnty. Sch. Bd., No. 23-170, slip op. at 1, 6–7 (U.S. Feb. 20, 2024) (Alito, J., dissenting from the denial of certiorari) (suggesting that a facially neutral public magnet school admission policy would be unconstitutional where it was designed to increase racial diversity).

159 *See generally* Cara McClellan, *When Claims Collide:* Students for Fair Admissions v. Harvard *and the Meaning of Discrimination*, 54 LOY. U. CHI. L.J. 953 (2023) (contrasting affirmative action with "mirror" claims of discrimination in the use of standardized test scores, *id.* at 954).

160 *Cf.* James Murphy & Virginia Carr Schneider, EDUC. REFORM NOW, A COMPELLING INTEREST 24–27 (2023), https://edreformnow.org/wp-content/uploads/2023/07/FINAL-Federal-RCA-SCOTUS-Report2.pdf [https://perma.cc/25X4-V4GZ] (detailing admissions reforms that should remain permissible post-*SFFA*).

Conclusion

The rapid rise of AI has been met with a raft of legislative proposals to target AI-enabled bias, many of which impose output-based requirements. The prevalence of testing and auditing mandates may be attributable in part to challenges unique to AI's mechanics. But given the chance to draft a new antidiscrimination regime on the blank slate of AI, legislators chose systemic remedies, targeted systemic harms, and rejected the notion that the entrenchment of bias in employment, healthcare, and housing[161] only matters when it results from intentional acts of discrimination. This Chapter explores whether the legal response to this new technology might also force a reflection about antidiscrimination law itself, rather than simply being a new technological "horse" onto which old regimes can be grafted.[162]

Refracting the AI legislative impulse onto legal regimes governing human decisionmakers is an opportunity to take stock of the antidiscrimination regimes we have inherited and consider what we want to carry forward. The prevalence of output-based regulations suggests a need to shore up the disparate impact doctrines that have fallen into legal uncertainty under the current Supreme Court. As both the *Meta* settlement and these regulations show, it is possible to provide robust, effects-based protections against AI-enabled discrimination without carving out an AI "exception." A narrow read of *Ricci* and *SFFA*, one sensitive to individual discrimination and which does not turn a blind eye to systemic bias, provides such a path forward. The nationwide response to AI points to a specific expectation of our institutions: a fair shake and the eradication of bias, whether intentional or not.

---

161 *See* Minow, *supra* note 53, at 171–72.

162 *See* The Law of the Horse (unpublished manuscript), *supra* note 14, at 46.