Stanford Law Review Online

Volume 79 September 2026

ESSAY

Unpacking AI Bias and the Antidiscrimination Law Dilemma

Hoang Pham, Hannah Cha & Rashon Poole*

Introduction

In July 2025, President Donald Trump issued an executive order titled "Preventing Woke AI in the Federal Government."[1] The order accuses artificial intelligence (AI) systems of having "ideological biases or social agendas" that distort accuracy and reliability, and directs federal agencies to only contract with developers of "unbiased" large language models. [2] Subsequent implementation guidance from the Office of Management and Budget, along with proposed federal legislation, litigation challenging the Colorado AI Act, and a proposed policy statement from the Federal Trade Commission, all suggest that disputes over AI bias may increasingly shape law—particularly antidiscrimination law—and public policy.[3]

* Hoang Pham is Director of Education and Opportunity at the Stanford Center for Racial Justice at Stanford Law School. Hannah Cha is an M.S. candidate in Computer Science at Stanford University and a Research Assistant at Microsoft Research. Rashon Poole is a Ph.D. candidate in Computer Science at the University of Michigan. This Essay is not possible without Ralph Richard Banks, whose steadfast guidance significantly challenged and influenced our ideas—for which we are beyond grateful. We owe many thanks to Ria Ellendula for her excellent research assistance. We are also deeply grateful to the editors of the *Stanford Law Review Online*, especially Hannah Dahleen and Faraaz Godil, for their careful edits and thoughtful suggestions that meaningfully improved the Essay.

1. Exec. Order No. 14319, 90 Fed. Reg. 35389 (July 28, 2025).
2. *Id.* at 35389-90.
3. *Overview of Administration's Policies on "Woke AI"*, FED'N ASS'NS IN BEHAV. & BRAIN SCIS. (Jan. 14, 2026), https://perma.cc/SX7K-4SGX; Press Release, Sen. Marsha Blackburn, *Blackburn Releases Discussion Draft of National Policy Framework for Artificial Intelligence* (Mar. 18, 2026), https://perma.cc/6MA8-J87H; Press Release, U.S. Dep't of Just., *Justice Department Intervenes in xAI Lawsuit Challenging Colorado's 'Algorithmic Discrimination' Law*, (Apr. 24, 2026), https://perma.cc/BG2R-9FJ8; FED. TRADE COMM'N, PROPOSED POLICY STATEMENT CONCERNING THE SUPPRESSION OF ACCURACY IN ARTIFICIAL INTELLIGENCE SYSTEMS 1 (2026), https://perma.cc/9TPH-ZVLH.

Critics have warned that the order raises civil rights concerns, creates vague compliance standards for technology companies, and overlooks the difficulty of developing a perfectly "unbiased" model.[4] Yet underlying the political rhetoric lies a more fundamental problem: Stakeholders define AI bias differently, depending on their values and the contexts in which AI is deployed. There is no clear single framework for navigating this complicated legal and technological landscape. [5] A lawyer may use the term "AI bias" to refer to racially discriminatory outcomes in a model designed to hire new employees; a developer may use it to describe statistical gender disparities in a model analyzing chest X-rays; and the Trump administration may define it as ideological slant when a model generates an image of ethnically diverse Vikings.[6] A survey of 146 papers analyzing bias in natural language processing systems found that a "majority of them fail to engage critically with what constitutes 'bias' in the first place." [7] Without recognizing that there are a multiplicity of values and domains in which AI bias can be defined, policymakers seeking to eliminate bias may be working toward an incoherent regulatory goal. What one framework or commentator identifies as a bias that requires intervention, another may view as a desirable feature of the AI model.[8]

Given that "different disciplines and communities understand [AI bias] differently,"[9] attempting to develop one definition for the field may be both unrealistic and counterproductive. This Essay argues that many contemporary disputes over AI bias are best understood as normative disputes over representation, which inherently implicates antidiscrimination legal principles.

---

4. *See, e.g.*, Tori Noble & Kit Walsh, *President Trump's War on "Woke AI" Is a Civil Liberties Nightmare*, ELEC. FRONTIER FOUND. (Aug. 14, 2025), https://perma.cc/NRJ6-NPDC (noting that government pressure on commercial companies and developers to modify model features may decrease the accuracy of AI models while increasing their likelihood to cause harm); Amos Toh, *How Trump's AI Policy Could Compromise the Technology*, BRENNAN CTR. FOR JUST. (Aug. 1, 2025), https://perma.cc/66EG-BM8Z (suggesting that the Trump Administration's executive order is too vague and requires developers to align with the administration's contested version of political reality); *Donald Trump is Waging War on Woke AI*, ECONOMIST (Aug. 28, 2025), https://perma.cc/22BD-4V93 (suggesting that building unbiased AI may be impossible because models are "black boxes," making it difficult to understand why they produce the responses they do and because developers must deal with philosophical questions for which there are no clear answers).

5. Ivan Kekez, Lode Lauwaert & Nina Begičević Redep, *Is Artificial Intelligence (AI) Research Biased and Conceptually Vague? A Systematic Review of Research on Bias and Discrimination in the Context of Using AI in Human Resource Management*, 81 TECH. SOC'Y art. 102818, at 2 (2025), https://perma.cc/YQ84-NZ9A.

6. *See, e.g.*, Exec. Order No. 14319, *supra* note 1, at 35389.

7. Su Lin Blodgett, Solon Barocas, Hal Daumé III & Hanna Wallach, *Language (Technology) is Power: A Critical Survey of "Bias" in NLP*, 58 PROCS. ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 5454, 5454 (2020), https://perma.cc/JJ4F-NJDF.

8. Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, 2018 PROCS. ACM/IEEE INT'L WORKSHOP ON SOFTWARE FAIRNESS 1, 7, https://perma.cc/LHA3-52N9.

9. SOLON BAROCAS, MORITZ HARDT & ARVIND NARAYANAN, FAIRNESS AND MACHINE LEARNING: LIMITATIONS AND OPPORTUNITIES 4 (2023).

Commentators bring to the bias debate different conceptions and aspirations for AI. Simply put, should AI systems reflect the world as it is—including its demographic patterns and historical inequalities—or should they be designed to promote a less discriminatory and more inclusive future? Should AI perpetuate existing patterns or aim to transform them? Accusations of AI bias over hiring or image generation tools often reveal more disagreement about what algorithms should represent rather than technical issues with the algorithm itself.

This representation dispute often arises from design choices about what a system should depict, prioritize, or optimize for—which inevitably involves tradeoffs between competing values like accuracy and fairness.[10] For example, an AI hiring tool for public school teachers trained on historical data may favor women candidates and be criticized as biased against men, yet it may also be defended as accurately reflecting past outcomes where a majority of public school teachers are women.[11] Should the algorithm be retrained to be more favorable to men, and if so, to what degree? Or should the algorithm—and other algorithms like it—represent historical trends as they were, even if it means there are disparate outcomes?

While this Essay focuses on these normative questions around AI bias, we acknowledge that technical issues are important as well, including challenges with incomplete or outdated training data that may lead to inaccurately skewed results.[12] The term "representation" is also used to describe some of these related problems. For example, existing work on fairness in machine learning discusses "representation bias," which arises "from how we sample from a population during [the] data collection process" leading to "[n]on-representative samples lack[ing] the diversity of the population."[13] There is additionally the idea of "representational harm," which occurs when "systems reinforce the

10. *See* Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 8 INNOVATIONS THEORETICAL COMPUT. SCI. CONF. 43:1, 43:5 (2017), https://perma.cc/D4PC-37KZ.

11. *See* NAT'L CTR. FOR EDUC. STATISTICS, CHARACTERISTICS OF PUBLIC SCHOOL TEACHERS 1 fig. 1 (2023), https://perma.cc/GS5J-4BS5 (documenting that 77 percent of public school teachers were female, while 23 percent were male).

12. The absence of data (whether fragmented, outdated, low quality, or missing) has been referred to as "algorithmic exclusion," which describes an AI system failing because it "lacks enough data on an individual [or group] to return an output about them." CATHERINE TUCKER, ARTIFICIAL INTELLIGENCE AND ALGORITHMIC EXCLUSION 2 (2025), https://perma.cc/B63J-9S4H. For example, Tiberius—the U.S. COVID-19 vaccine allocation algorithm—used data from the American Community Survey of 2018 to allocate vaccines when supply was scarce. However, an analysis of the 2020 United States Census found that certain groups were undercounted in that survey, including Black Americans by 3.3 percent. *Id.* at 4. This incomplete data can produce inaccurate and skewed algorithmic outcomes favoring some groups over others.

13. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman & Aram Galstyan, *A Survey on Bias and Fairness in Machine Learning*, 54 ACM COMPUTING SURVS. art. 115, at 5, (2021), https://perma.cc/RL2P-FYW6.

subordination of some groups along the lines of identity."[14] This type of harm most commonly occurs through the perpetuation of stereotypes, such as search engine results that reinforce racial stereotypes.[15] Our understanding of the representation dispute builds on these concepts by showing there is a contested, threshold normative question to also explore: What should an algorithm *even represent*? The representation dispute *informs* whether and how we address problems with data or the consequences that may result from using that data.

Further, we acknowledge there are other related AI challenges—such as privacy and surveillance, transparency, and procedural fairness—that might not map neatly onto our representation framing. There are also other forms of bias, like political bias, which do not necessarily focus on *group* representation like we do in this Essay. Nevertheless, when these issues do implicate bias, we suggest that similar questions about representation can help stakeholders navigate both what "bias" means within a given context and perhaps what to do about it.[16]

Part I of this Essay canvasses the various definitions of AI bias, showing that although these different conceptions may illuminate important nuances, they also make understanding and addressing AI bias more difficult. Part II examines the representation dispute at the core of AI bias, demonstrating how these disputes vary depending on the context of model deployment and typically involve moral and legal questions around historical data as well as model outputs. Part III explores the implications of the representation dispute for antidiscrimination law and AI, analyzing an evolving legal framework that is increasingly at odds with rapid technological advancement—what we consider the antidiscrimination law dilemma. Finally, Part IV proposes an inquiry-based, context-specific approach to unpacking AI bias that can help stakeholders navigate the uncertain landscape of law and policy in the AI era.

I. Conceptions of AI Bias

It is tempting to say that bias in computing systems is a recent phenomenon; however, the conversation regarding fairness in this context has been active for decades. While generative models and large-scale algorithmic decisionmaking systems have brought the issue to public attention, earlier scholars such as James H. Moor, Deborah G. Johnson, and John M. Mulvey were already examining

14. BAROCAS ET AL., *supra* note 9, at 251.

15. *See generally* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 1-14 (2018).

16. For example, political bias often comes down to which political views are being represented by a model—which is, at its core, still a dispute over representation. When testing AI models for political bias, researchers found that models trained on the internet were biased toward conservative views, whereas new models often trained on curated human feedback—that is, curating what they want the model to *represent*—were more biased toward liberal views. Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang & Tatsunori Hashimoto, *Whose Opinions Do Language Models Reflect?* 2023 PROCS. INT'L CONF. ON MACHINE LEARNING 29971, 29971, https://perma.cc/R9JG-PVY6.

ethical concerns arising from "computer decision systems."[17] Building on that scholarship, Batya Friedman and Helen Nissenbaum's seminal 1996 paper, *Bias in Computing Systems*, provided a detailed framework for understanding how bias can arise.[18] Today, literature across fields has most commonly defined AI bias as algorithmic skewing that produces unfair or discriminatory outcomes toward an individual or group.[19] But when, exactly, does algorithmic skewing produce unfair or discriminatory outcomes? One common answer is when it violates antidiscrimination law, particularly when its usage produces a disparate impact by disproportionately harming certain groups based on legally protected characteristics such as race or sex, without sufficient justification. Yet despite this long history of study and some commonalities, there still exists a variety of ways to define AI bias, whether in the technical or regulatory realm.

Among computer scientists and technical researchers, the definition of bias remains a subject of debate. One common perspective takes a mathematical approach, where a biased system is one whose decisions are unfairly skewed statistically toward a particular group of people.[20] However, other researchers view bias as a sociotechnical problem, where simply equalizing outcomes across groups fails to account for the context in which a system is deployed.[21] From this perspective, a model that treats all groups "identically" might be considered biased if it ignores the deployment context or fails to account for the historical inequalities that may be embedded in its training data. These are only a few among many definitions of bias, all of which cannot be reconciled simultaneously, leaving the determination of what is "biased" to be worked out between a variety of stakeholders with different values and aspirations for AI.

Recent legislation reflects more overlap, generally moving closer to a sociotechnical and context-aware approach than to a strict rule of identical treatment or equalized outcomes. For example, before it was repealed and replaced, the Colorado AI Act defined "algorithmic discrimination" as "any

17. *See* James H. Moor, *What Is Computer Ethics?*, 16 METAPHILOSOPHY 266, 266-68 (1985); Deborah G. Johnson & John M. Mulvey, *Accountability and Computer Decision Systems*, 38 COMMC'NS ACM 58, 58-59 (1995).

18. Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM TRANS. INFO. SYS., 330, 332-33 (1996).

19. *See, e.g.*, Richard Ribón Fletcher, Audace Nakeshimana & Olusubomi Olubeko, *Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health*, 3 FRONTIERS A.I. art. 561802, at 6 (2021), https://perma.cc/VXA3-JVMS (defining bias as "a systematic error or an unexpected tendency to favor one outcome over another," including "when an algorithm has an undesired dependence on a specific attribute in the data that can be attributed to a demographic group").

20. *See, e.g.*, Mehrabi et al., *supra* note 13, at 2 (defining an "unfair algorithm" as "one whose decisions are skewed towards a particular group of people").

21. *See, e.g.*, Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian & Janet Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, 2019 PROCS. CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 59, 59, https://perma.cc/4H62-YQ2R (arguing that fairness and justice do not have a meaningful application to technology when divorced from the social context in which technology systems are deployed).

condition in which the use of an AI system results in an unlawful differential treatment or impact that disfavors an individual or group of individuals on the basis" of a legally protected characteristic, such as race, sex, religion, or disability.[22] Importantly, the Act did not prohibit developers or deployers from using an AI system to expand an "applicant, customer, or participant pool to increase diversity or redress historical discrimination."[23] Here, although bias is not explicitly defined, it is seemingly built into the law's definition of "algorithmic discrimination."

In another example, New York City's Local Law 144 requires employers to conduct an independent bias audit before using an "automated employment decision tool," and defines bias through a mathematical equation: the Equal Employment Opportunity Commission's "four-fifths" rule for disparate impact.[24] The audit requires the employer to calculate the selection rate (i.e., the proportion of candidates from a specified demographic that are selected), and based on this, the "impact ratio."[25] More generally, impact ratios are used to determine whether outcomes between groups differ in a meaningful way, and are calculated by taking the highest selection rate that exists across groups and comparing it to the selection rate of another (e.g., if the highest selection rate is 48 percent, that group's impact ratio is 1.0, and a group with a selection rate of 47 percent would be 0.979).[26] An impact ratio that falls below 0.8 (i.e., four-fifths) could increase suspicion of disparate impact.[27]

Practical challenges arise due to these nuances in defining AI bias. How do policymakers create legislation that broadly governs AI, whether across a state or the nation? How might lawyers bring legal claims presented in different contexts, such as employment, education, housing, healthcare, and consumer finance? How should developers design systems to be legally compliant if "bias" in the regulatory sense does not match their own technical definition? At the same time, stakeholders bring different values to the debate and the many contexts in which algorithms are deployed may also entail different definitions. Therefore, while this Essay does not attempt to state a universal definition for the field, it does argue that one approach to navigating the various conceptions of AI bias is to unpack and examine it as a fundamental dispute over representation.

---

22. COLO. REV. STAT. § 6-1-1701(1)(a) (2024).

23. COLO. REV. STAT. § 6-1-1701(1)(b)(I)(B) (2024).

24. NYC DEP'T OF CONSUMER & WORKER PROT., *Automated Employment Decision Tools: Frequently Asked Questions* 1-2 (2023), https://perma.cc/TE53-4AST; 29 C.F.R. § 1607.4(D) (2026).

25. 6 R.C.N.Y. § 5-300 (2023) (defining "Impact Ratio").

26. *Id.*

27. Lucas Wright et al., *Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability*, 2024 ACM CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 1701, 1708, 1713, https://perma.cc/FGA7-QJPZ.

II. Algorithms and the Representation Dispute

There are countless decisions made when technologists design algorithmic tools, one of the most significant being what a system's output should be, including what it should depict, prioritize, or optimize for. A central question around system outputs involves representation: should AI systems reflect the world as it is—including demographic patterns and historical inequalities—or should they be designed to promote a less discriminatory and more inclusive future? AI bias implicates normative disputes over representation. These disputes can be broken into three categories: (1) using historical data to represent the past, but that representation itself may be unfair or discriminatory; (2) using historical data to represent the past, but that representation can be helpful in improving the human condition; and (3) steering AI to advance representational goals.

A. Representing an Unfair or Discriminatory Past

The representation dispute is best understood by first examining AI bias as it is commonly defined: algorithmic skewing that produces unfair or discriminatory outcomes toward an individual or group. Consider the college admissions context. Research has shown that algorithms trained on historical student data to predict college success can produce outcomes that favor White and Asian students.[28] Nonetheless, a university may use these types of algorithms in the admissions context to assist with decisionmaking on tens of thousands of applications. Here, using historical data may be important because it helps the university select qualified students—that is, those who are most likely to be successful in college—which is precisely what an admissions tool would be intended to do.

Others, however, might argue that the algorithm should be designed to be more representative of historically underrepresented groups, to promote not only a more inclusive future but greater algorithmic accuracy that aligns outcomes with the tool's intended purpose. During training, for example, the model may learn "spurious correlations," a technological phenomenon occurring when the model relies on non-predictive attributes that happen to be correlated with the target variable—here, college success—which may reinforce disparate outcomes.[29] In the context of college success, non-predictive attributes may include variables such as demographic characteristics (e.g., race) and socioeconomic traits (e.g., receiving free or reduced lunch). A university may seek to identify qualified applicants with the assistance of this AI tool, but

28. Denisa Gándara, Hadis Anahideh, Matthew P. Ison & Lorenzo Picchiarini, *Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction*, 10 AERA OPEN art. 23328584241258741, at 7-9 (2024), https://perma.cc/3KVH-M32L.

29. *See* Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu & Aidong Zhang, Spurious Correlations in Machine Learning: A Survey 1-2, 7 (Feb. 20, 2024) (unpublished manuscript), https://perma.cc/5LGX-H9H5.

because the tool is trained on a variety of datapoints from past student applications, the model may learn to correlate data associated with students in the minority with a lower likelihood of college success.

Indeed, research has found that these models are "more likely to predict failure for students who actually succeed if those students are categorized as Black or Hispanic."[30] Even though these correlations may mirror historical data and seem to perform accurately during model training, their use in actual decisionmaking can be unfair or discriminatory if they disproportionately disadvantage qualified students based on proxies for race rather than criteria that meaningfully reflect academic preparation or potential. This is known as "target specification bias," where the way a goal or target is operationalized (e.g., using broad historical student data to train an algorithm) does not match the objective defined by decisionmakers (e.g., identify qualified applicants by predicting college success).[31]

These issues exist in a variety of sectors deploying AI, and they raise legal implications as well as broader moral and ethical ones.[32] What should be the normative goals of an algorithm? Who gets to decide those normative goals? Understanding bias as a fundamental dispute over representation recognizes that these questions will inevitably be answered differently depending on who, why, what kind, and where AI is being deployed.

B. Representation to Improve the Human Condition

Although algorithmic fairness literature primarily discusses AI bias as a negative phenomenon, not all disparate outcomes warrant correction. Instead, in some contexts, using purely historical data to represent the past may be helpful in improving the human condition.[33] For example, risk assessment tools in healthcare may account for race, age, or sex to accurately assess disease outcomes. Even when these algorithmic predictions may skew disproportionately toward a legally protected group—for example, mortality

30. Gándara et al., *supra* note 28, at 2.

31. Eran Tal, *Target Specification Bias, Counterfactual Prediction, and Algorithmic Fairness in Healthcare*, 2023 PROCS. AAAI/ACM CONF. AI, ETHICS, & SOC'Y 312, 312-13, https://perma.cc/9SUA-MTRF.

32. *See, e.g.*, Doe 1, et al. v. Meta Platforms, Inc., No. 26-cv-07122, 2026 WL 2076139 (N.D. Cal. July 17, 2026) (employment termination); Brief of the Equal Employment Opportunity Commission as Amicus Curiae in Support of Plaintiff at 1, Mobley v. Workday, Inc., 750 F. Supp. 3d 796 (N.D. Cal. July 12, 2024) (No. 23-cv-00770), ECF No. 60-1 (employment hiring); Huskey v. State Farm Fire & Cas. Co., No. 22-cv-07014, 2025 WL 3552388, at *1 (N.D. Ill. Dec. 11, 2025) (insurance); Louis v. SafeRent Sols., LLC, 685 F. Supp. 3d 19, 25 (D. Mass. 2023) (housing).

33. Mirjam Pot, Nathalie Kieusseyan & Barbara Prainsack, *Not All Biases Are Bad: Equitable and Inequitable Biases in Machine Learning and Radiology*, 12 INSIGHTS INTO IMAGING art. 13, at 8 (2021), https://perma.cc/9FDP-V2LF.

rates for heart disease and stroke are highest among African Americans[34]—this algorithmic skewing or form of differentiation reflects statistically grounded variation and can enhance the effectiveness of systems without producing unfair or discriminatory outcomes. In fact, it may "enable targeted interventions and preventative measures, thereby fostering greater equity in healthcare."[35]

This "social good" function of AI cannot be ignored.[36] While efforts to address AI bias usually focus on remedying harm, it is increasingly important to recognize how AI can address some of society's most entrenched inequalities. For example, an AI-powered flood forecasting tool may produce suggestions or outputs that skew toward supporting low-income Black communities but are nonetheless representative of existing data, as the Congressional Budget Office estimates that low-income Black households "face the largest increases in flood risk from 2020 to 2050."[37] In this context, the disproportionate representation of low-income Black communities may actually improve their life outcomes if government agencies use the tool to support those who face the greatest flood risks. Notably, research suggests to "protect at-risk populations and improve the overall accuracy of [flood risk management] plans, data-driven projections must be equitable and inclusive."[38]

It is still important to identify harms that may arise in these contexts and appropriate mitigation strategies so that AI's benefits can be fully realized.[39] The challenge, then, is to determine when representational goals reflect acceptable or even beneficial outcomes that are not discriminatory. That determination cannot be made purely on technical grounds, such as whether an algorithmic outcome skews toward a particular group, but it requires normative judgment about the risks, context, and purposes for which AI systems are used.

34. George A. Mensah, *Cardiovascular Diseases in African Americans: Fostering Community Partnerships to Stem the Tide*, 72 AM. J. KIDNEY DISEASES S37, S37 (2018).

35. Daniel Amponsah, Ritu Thamman, Eric Brandt, Cornelius James, Kayte Spector-Bagdady & Celina M. Yong, *Artificial Intelligence to Promote Racial and Ethnic Cardiovascular Health Equity*, 18 CURRENT CARDIOVASCULAR RISK REPS. 153, 155 (2024), https://perma.cc/ZUZ9-KXBS.

36. *See generally* Nenad Tomašev et al., *AI for Social Good: Unlocking the Opportunity for Positive Impact*, 11 NATURE COMMC'NS. art. 2468 (2020).

37. CONG. BUDGET OFF., *Communities at Risk of Flooding* (Sept. 2023), https://perma.cc/2AC8-KF7T.

38. Peigen Wang, Xiaoxu Wu & Yichen, *AI-Driven Approaches to Flood Risk Management: Overcoming Data Bias and Enhancing Decision-Making*, 50 CLIMATE RISK MGMT. 1, 14 (2025), https://perma.cc/82EL-HVKN.

39. *See, e.g.*, Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 453 (Oct. 25, 2019), https://perma.cc/A9LA-L7JK (finding that reformulating a health care intervention algorithm to stop using health costs as a proxy for health needs eliminated racial discrepancy in predictions).

C. Steering AI to Advance Representational Goals

Rather than relying on historical data as it exists, AI can also be steered to advance representational goals. "Steering" an AI system refers to the process of guiding or influencing the model's behavior to produce outputs that are aligned with desired goals, values, or contexts.[40] Representational goals, such as increasing the number of women on corporate boards, are often determined in response to social inequality and aim to strike greater balance between different groups. For example, efforts to steer AI to advance representational goals may include creating models that generate more gender-balanced depictions of occupations typically dominated by certain genders.[41]

Nevertheless, the "woke AI" executive order would likely consider an AI system steered to advance representational goals as "ideological bias."[42] The order suggests these biases result from the inclusion of "destructive" ideologies like diversity, equity, and inclusion (DEI) in AI models, producing distorted outputs.[43] For example, the order cites Gemini, Google's AI assistant, which inaccurately "changed the race or sex of historical figures—including the Pope, the Founding Fathers, and Vikings—when prompted for images."[44] Critics, including the current administration, may argue that some forms of AI bias exist not because of historical data but rather due to "manipulation of racial or sexual representation in model outputs . . . [that] displaces the commitment to truth in favor of preferred outcomes."[45]

Indeed, in developing Gemini, Google attempted to "remove biases" by increasing the representation of "diverse people" in Gemini's outputs—seeking to balance how individuals were depicted.[46] However, using this "balancing" strategy as a blanket solution in *all* generative imagery can result in altering the race or sex of historically specific figures. Designing a model to increase representation in contexts where a particular group is underrepresented, such as various occupations, can promote fairness and inclusivity. But altering

40. Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adserà & Mikhail Belkin, *Toward Universal Steering and Monitoring of AI Models*, 391 SCIENCE 787, 787 (2026).

41. *See, e.g.*, Luhang Sun et al., *Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image-Generative AI*, 29 J. COMPUT.-MEDIATED COMMC'N art. Zmad045, at 1-2 (2024), https://perma.cc/XX5D-ZV5K (finding that generative image models perpetuate occupational gender biases in depictions of men and women in various occupations).

42. *See* Exec. Order No. 14319, *supra* note 1, at 35389-90.

43. *Id.* at 35389.

44. *Id.*; *see also Is Google's Gemini Chatbot Woke by Accident, or by Design?*, ECONOMIST (Feb. 28, 2024), https://perma.cc/64FN-TDYK (describing Gemini generating historically inaccurate images of the Vikings, the Pope, the Nazis, and the Founding Fathers).

45. Exec. Order No. 14319, *supra* note 1, at 35389.

46. Sarah Shamim, *Why Google's AI Tool Was Slammed for Showing Images of People of Colour*, AL JAZEERA (Mar. 9, 2024), https://perma.cc/JV4D-998J.

historical representations altogether can be seen as undermining the truth and contributing to public distrust in AI systems.[47]

This raises a key question in the representation dispute: Should AI systems be designed to promote a less discriminatory and more inclusive future? It depends on who you ask. While the creation of images that attempt to be more balanced of underrepresented groups generally falls legally under the First Amendment[48]—which we do not analyze here—contexts where an AI system appears to unfairly withhold services, resources, or opportunities constitute allocative harms and implicate antidiscrimination legal principles.[49] Some would argue that allocative harms resulting in unequal outcomes should be remedied through less discriminatory algorithms.[50] Another view, reflected in the executive order and by the Trump administration, might suggest that algorithms designed to "advance 'diversity' or 'redress historic discrimination'" are considered "woke DEI ideology" and are illegal.[51]

Yet the President's prohibition on "woke AI" may actually violate antidiscrimination law, which sometimes requires attention to group differences.[52] In other words, what is labeled as "ideological bias" may in some cases reflect attempts to comply with, rather than depart from, existing legal norms, including the Civil Rights Act of 1964, the Fair Housing Act of 1968, and the Equal Credit Opportunity Act of 1974.[53] Nonetheless, the future of antidiscrimination law in the AI era is unsettled, and using these established

47. *See* Cynthia Dwork & Martha Minow, *Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law*, 151 DÆDALUS, Spring 2022, at 309, 310.

48. *See* Eugene Volokh, Mark A. Lemley & Peter Henderson, *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 651, 652 (2023).

49. BAROCAS ET AL., *supra* note 9, at 19.

50. *See, e.g.*, Talia Gillis, Vitaly Meursault, & Berk Ustun, *Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending*, 2024 PROCS. ACM CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 377, 377, https://perma.cc/3L53-TTYH (proposing a formal audit method to search for less discriminatory models—particularly in the fair lending context—that reduces disparate impact without compromising business performance).

51. U.S. Dep't of Just., *supra* note 3.

52. *See, e.g.*, United Steelworkers of Am. v. Weber, 443 U.S. 193, 208-09 (1979) (holding that Title VII of the 1964 Civil Rights Act does not prohibit private sector employers to voluntarily adopt race-conscious affirmative action programs "designed to eliminate conspicuous racial imbalance in traditionally segregated job categories," so long as it does not "unnecessarily trammel the interests of the white employees").

53. Civil Rights Act of 1964, Pub. L. No. 88-352, 78 Stat. 241 (codified as amended in scattered sections of the U.S. Code); Fair Housing Act of 1968, Pub. L. No. 90-284, 82 Stat. 73 (codified as amended at 42 U.S.C. §§ 3601-3619, 3631); Equal Credit Opportunity Act of 1974, Pub. L. No. 93-495, tit. V, 88 Stat. 1500, 1521-25 (codified as amended at 15 U.S.C. § 1691, 1691a-1691c, 1691d-1691e).

legal principles to promote a less discriminatory or more inclusive future may be increasingly difficult to do.[54]

III. The Antidiscrimination Law and AI Dilemma

Antidiscrimination legal doctrines have historically developed in contexts involving human decisionmakers, where questions of intent and causation are already complex. The rapid advancement of AI technology—trained on massive amounts of data that make it nearly impossible to determine how an AI system made its decisions (i.e., the "black box" problem)—complicates these inquiries even more.[55] In particular, longstanding legal frameworks that have recently leaned further into emphasizing intent and anticlassification, especially with regards to race discrimination under the Equal Protection Clause, may be progressively inadequate to evaluate systems that produce disparities without explicit intent or clear categorical distinctions.[56]

As a result, even where there is agreement about the representation dispute—perhaps that AI should be designed to promote a less discriminatory future—there is less consensus about how existing antidiscrimination law would apply and whether new frameworks are needed. Recent interdisciplinary discussions have emphasized that advances in AI are likely to shift antidiscrimination law's focus toward disparate impact analysis, while also raising challenging questions around a variety of related issues, including who should be liable for algorithmic discrimination (e.g., developers, deployers, or both) and what type of evidence is needed to establish a claim.[57] But even the inclination to focus more on disparate impact is complicated by the Supreme

54. *See, e.g.*, *Developments in the Law—Resetting Antidiscrimination Law in the Age of AI*, 138 HARV. L. REV., 1562, 1584 (2025) (arguing that existing antidiscrimination frameworks leave gaps when applied to AI-enabled decisionmaking).

55. Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889, 894 (2018).

56. *See, e.g.*, Washington v. Davis, 426 U.S. 229, 239-42 (1976) (holding that absent proof of discriminatory intent, disparate impact alone does not establish an Equal Protection Clause violation); Students for Fair Admissions, Inc. v. President & Fellows of Harv. Coll., 600 U.S. 181, 217-18 (2023) (striking down race-based affirmative action in college admissions, further narrowing the scope for when racial classifications are permissible); Louisiana v. Callais, 146 S. Ct. 1131, 1156, 1162 (2026) (holding that drawing legislative districts based on race violates the Equal Protection Clause unless there exists a compelling interest, such as complying with Section 2 of the Voting Rights Act, which requires "a strong inference that intentional discrimination occurred").

57. *See* Monica Schreiber, *Conference Addresses Impact of AI on Antidiscrimination Law*, STAN. REP. (Mar. 19, 2026), https://perma.cc/2L3J-EGX7; Chiraag Bains, *The Legal Doctrine That Will Be Key to Preventing AI Discrimination*, BROOKINGS INST. (Sept. 13, 2024), https://perma.cc/NS7T-V4F3.

Court's move toward a colorblind constitution,[58] inconsistent government enforcement,[59] and a constantly evolving technological landscape.

Consider current disparate impact doctrine under various civil rights statutes, which generally involves a three-step analysis.[60] First, a plaintiff must show that an algorithm had a disparate impact on a protected group. Second, the defendant has the burden of showing a legitimate business justification for the algorithm it deployed. Third, even if the defendant could show a legitimate business justification, it can still be liable if the plaintiff can show "an alternative that would serve the same ends with less disparate impact"—what some refer to in the algorithmic context as a "less discriminatory algorithm" (LDA).[61] Whether based on a legal duty to search for an LDA or an interest to affirmatively protect against liability, this disparate impact framework would theoretically prompt companies that deploy AI in high-risk contexts such as employment to make sure they deploy an LDA if their algorithm does produce discriminatory outcomes.

However, would the process of searching for an LDA be in and of itself discriminatory if it required a developer to take into account race or ethnicity in its model design? What if the development process only involved an

58. While *Griggs v. Duke Power Co.* established disparate impact liability under Title VII of the Civil Rights Act of 1964—holding that facially neutral employment practices that are "discriminatory in operation" are prohibited unless the practice is related to job performance—it has always run counter to *Washington v. Davis*, which has long required proof of discriminatory intent under the Equal Protection Clause. *See* Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971); *Davis*, 426 U.S. at 240. Justice Scalia's concurrence in *Ricci v. DeStefano* squarely addresses this tension, suggesting that it will be an "evil day on which the Court will have to confront the question: Whether, or to what extent, are the disparate-impact provisions of Title VII of the Civil Rights Act of 1964 consistent with the Constitution's guarantee of equal protection?" Ricci v. DeStefano, 557 U.S. 557, 594 (2009) (Scalia, J., concurring). Continuing, he says, "if the Federal Government is prohibited from discriminating on the basis of race, then surely it is also prohibited from enacting laws mandating that third parties—e.g., employers, whether private, State, or municipal—discriminate on the basis of race." *Id.* Justice Scalia's reasoning is only strengthened by the Supreme Court's recent decisions furthering a colorblind constitution. *See Students for Fair Admissions*, 600 U.S. at 217-18; *Callais*, 146 S. Ct. at 1156, 1162.

59. *See, e.g.*, Ralph Richard Banks, *Reassessing Disparate Impact*, STAN. CTR. FOR RACIAL JUST. (Oct. 2, 2025), https://perma.cc/GZ5C-J8QR (arguing that while the Trump Administration's "Restoring Equality of Opportunity and Meritocracy" executive order, Exec. Order No. 14281, 90 Fed. Reg. 17,537 (Apr. 23, 2025), rejected the legal theory of disparate impact, the Administration has also invoked the order in calling for universities to make available their hiring and admissions data to investigate discrimination); Constitutionality of Disparate-Impact Liability Under Title VII, 50 Op. O.L.C. (June 9, 2026) (slip op. at 1-2) (addressing the Chair of the U.S. Equal Employment Opportunity Commission and explaining that the U.S. Department of Justice now interprets disparate impact liability under Title VII as proscribing "only those practices that reflect a significant likelihood of *intentional* discrimination," claiming that previous interpretations that "contemplate[d] liability based on disproportionately adverse effects alone" are unconstitutional (emphasis added)).

60. Emily Black, John Logan Koepke, Pauline T. Kim, Solon Barocas & Mingwei Hsu, *Less Discriminatory Algorithms*, 113 GEO. L.J. 53, 59 (2024).

61. *Id.* at 72.

"awareness of race . . . akin to decisions like where to locate a school"—might that still be considered discriminatory?[62] Some argue that "[t]he Supreme Court has repeatedly stated that one of Congress's purposes in passing civil rights laws was to spur 'self-examin[ation]' and 'self-evaluat[ion],' with the goal of eliminating arbitrary discriminatory practices. Recognizing that voluntary compliance is key, courts have approved proactive efforts to remove sources of bias."[63] Others have suggested that the Supreme Court in *Students for Fair Admissions v. President and Fellows of Harvard College* "left very little room for explicit race-conscious antidiscrimination interventions, potentially posing challenges for the algorithmic fairness community, whose work typically involves formalizing a fairness metric, constraint, or objective that is conscious of the protected attribute, with the goal of affirmatively changing the model to be 'fairer.'"[64] If the former is true, a developer could at the least address algorithmic discrimination by having an "awareness of race" or other protected characteristics to evaluate the unfairness of different models, ensuring an LDA is deployed. But if the latter is true, it is unclear whether a new legal framework or technological innovation would emerge to address disparate outcomes resulting from AI deployment.

Emily Black et al. propose a novel "model multiplicity" approach to searching for LDAs, which posits that there are multiple equally performing models that exist for the same prediction task, and "some interchangeable models will have less discriminatory effect."[65] Notably, the authors argue this approach "does not require the use of demographic data during model training. Instead, this data would only be used for testing and evaluation of varying possible models [i.e., an 'awareness of race']—methods that should not trigger disparate-treatment concerns."[66] However, we suggest it is uncertain whether using demographic data for testing and evaluation as well as other machine learning stages, like model selection, would run afoul of antidiscrimination law. Emily Black et al. cite Pauline Kim's foundational work on this issue to support their assertion.[67] There, Pauline Kim argues that these methods do not trigger disparate-treatment concerns because they are "more accurately understood as removing bias from processes that would otherwise be unfair." [68] In a hypothetical example, Pauline Kim suggests removing supervisor evaluations included in training data for an employment selection algorithm that

62. *Id.* at 119.

63. *Id.* at 116.

64. Alice Xiang, *Mirror, Mirror, on the Wall, Who's the Fairest of Them All?*, DÆDALUS, Winter 2024, at 250, 257.

65. *See* Black et al., *supra* note 60, at 62.

66. *Id.* at 71.

67. *Id.* (citing Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action*, 110 CALIF. L. REV. 1539, 1574-83 (2022)).

68. Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action*, 110 CALIF. L. REV. 1539, 1576-77 (2022).

"consistently downgraded Black employees relative to others even though they demonstrated the same level of productivity" illustrates non-discriminatory race-consciousness.[69]

But what if a model produces racially disparate outcomes that are not the result of unfairness (see Section II.B.)? Additionally, given the black box nature of algorithms, it seems that determining whether a disparate result was concretely due to unfairness may be challenging, particularly in systems with massive datasets like large language models. Yet model multiplicity literature often treats disparate outcomes as inherently unfair without addressing this ambiguity.[70] Doing so likely would subject these methods to disparate-treatment concerns precisely because it is difficult to pinpoint legitimate unfairness versus benign demographic differences based in training data.

Gordon Dai et al.'s important work on model multiplicity raises two related issues. First, it confirms the novelty of the model multiplicity approach, whereby "looking across many models selected on the basis of accuracy and searching for the fairest among them" corresponds to "an LDA search that does not directly use protected attribute information until after all the models are trained, i.e., only as a step to evaluate models and choose among them post-hoc."[71] While *Ricci v. DeStefano* is a case factually distinct from the algorithmic context, we are not convinced that "*nothing* in [*Ricci*] suggests that choosing a less discriminatory alternative among equally effective models prior to implementation is disparate treatment [emphasis added]," as Emily Black et al. argue.[72] In *Ricci*, the City of New Haven discarded promotion examination results that favored primarily White firefighters to avoid disparate impact liability.[73] The Court held that an employer may not discard examination results "to achieve a more desirable racial distribution of promotion-eligible candidates—absent a strong basis in evidence that the test was deficient and that discarding the results is necessary to avoid violating the disparate-impact provision."[74] If there is no "strong basis in evidence" that a model with more disparate impact in fact violates disparate impact law (which, as discussed above, may be challenging to do), it is plausible that "searching for [and choosing] the fairest among" multiple models would constitute prohibited disparate

69. *Id.* at 1576.

70. *See, e.g.*, Gordon Dai, Pavan Ravishankar, Rachel Yuan, Emily Black & Daniel B. Neill, *Be Intentional About Fairness!: Fairness, Size, and Multiplicity in the Rashomon Set*, 5 PROCS. ACM CONF. EQUITY & ACCESS IN ALGORITHMS, MECHANISMS, & OPTIMIZATION 42, 49 (2025), https://perma.cc/LE4U-M7HP (acknowledging that in one LDA search method, protected characteristics are only used as a "step to evaluate models and choose among them post-hoc"—without acknowledging whether this step directly addresses unfairness).

71. *Id.*

72. *See* Black et al., *supra* note 60, at 117

73. Ricci v. DeStefano, 557 U.S. 557, 562 (2009).

74. *Id.* at 584.

treatment—especially given the Supreme Court's active shift toward a colorblind constitution.[75]

Second, Gordon Dai et al. found that choosing a random model from a set of retrained models "will have an extremely low chance of being the fairest, or even one of the fairer, models within the set."[76] They suggest the optimal LDA search method "uses some direct minimization of disparities across demographic groups during the model creation process, whether that be in hyperparameter tuning, optimization, or other parts of the pipeline"—but acknowledge the "disagreement in the legal literature as to whether and to what extent interventions for disparity reduction across demographic groups that use protected class information are legally permissible."[77]

This uncertainty is further illustrated in a recent lawsuit brought by xAI challenging the Colorado AI Act, where the U.S. Department of Justice (DOJ) joined as an intervenor in April 2026. In the DOJ's Complaint in Intervention, the government argues that because the Colorado AI Act requires developers and deployers to prevent the risk of disparate outcomes based on demographic characteristics, it effectively requires them to "expressly use demographic characteristics, including race, sex, and religion when building and using algorithmic models"—which the DOJ suggests violates the Equal Protection Clause.[78] This reasoning is consistent with the DOJ's recent slip opinion explaining that the DOJ now interprets disparate impact liability under Title VII of the Civil Rights Act of 1964 as proscribing "only those practices that reflect a significant likelihood of intentional discrimination," claiming that previous interpretations that "contemplate[d] liability based on disproportionately adverse effects alone" are unconstitutional.[79] It also reflects the broader trend in recent Supreme Court precedent where the consideration of race—even to remedy past discrimination—is increasingly unable to satisfy strict scrutiny.[80] Notably, on the same day the DOJ intervened, xAI and the

75. *See supra* note 58.

76. Dai et al., *supra* note 70.

77. *Id.*

78. Complaint in Intervention ¶ 48, United States v. Weiser, No. 26-cv-01515 (D. Colo. Apr. 24, 2026), ECF No. 12-2.

79. Constitutionality of Disparate-Impact Liability Under Title VII, *supra* note 59.

80. *See, e.g.*, Students for Fair Admissions, Inc. v. President & Fellows of Harv. Coll., 600 U.S. 181, 214-18 (2023) (holding that the diversity rationale put forth by the defendant-universities to justify race-based affirmative action was not sufficient to satisfy strict scrutiny); Louisiana v. Callais, 146 S. Ct. 1131, 1160 (2026) (holding that the Fifteenth Amendment prohibits "present-day intentional racial discrimination," and that discrimination that "occurred sometime ago, as well as present-day disparities that are characterized as the ongoing 'effects of societal discrimination,' are entitled to much less weight"); Allen v. Milligan, 146 S. Ct. 1377, 1381, 1384 (2026) (staying a preliminary injunction preventing Alabama from using congressional district maps that the District Court found intentionally discriminated, even after *Callais*, citing failure to heed the "presumption of legislative good faith"). However, the *Allen* decision suggests that the

footnote continued on next page

Colorado Attorney General filed a joint motion to suspend enforcement of the law, and in May 2026, the Colorado Governor signed a replacement law that makes no mention of protected characteristics, among other significant changes.[81]

These disagreements and developments suggest that the relationship between antidiscrimination law and AI needs to evolve—particularly with sustained collaboration between technical, legal, and policy communities.

IV. An Inquiry-Based, Context-Specific Approach to AI Bias

As AI becomes mass-deployed across society, each sector presents a variety of applications where accusations of bias may arise. Unpacking bias through the representation dispute shows how asking foundational questions about the normative goals of AI systems can help us better understand the values and contexts in which bias can be defined—and what to do about it. Instead of presuming every algorithmic outcome that produces a disparate impact on a protected group is AI bias, a more nuanced approach might prioritize inquiry over generalizing—supporting the development of reasoned remedies that fit the context in which a system is deployed.[82] An ideal set of questions for stakeholders could be:

1. Is a given algorithmic dispute one about representation (i.e., reflecting the world as it is or altering it to promote a less discriminatory and more inclusive future)? If so, what values are implicated? Which context is AI being deployed in?

2. What type of impact does the algorithm have on a legally protected group, if any at all? If there is an impact, which laws are implicated and how do those laws shape whether a remedy is needed?

3. If a remedy is needed, what interdisciplinary solutions would be responsive to the harm, legally compliant, and technologically feasible?

On the one hand, an algorithm that represents historical data but has a positive impact on a legally protected group may require monitoring, such as

"strong inference of intentional discrimination" standard set in *Callais* may also be exceedingly difficult to meet. *Id.* at 1384; *Callais*, 146 S. Ct. at 1163.

81. *See* Marc B. Collier, Helen Christakos, Ethan Glenn & Shushan Gabrielyan, *X.AI Sues, DOJ Intervenes, Enforcement of Colorado's AI Act Suspended*, NORTON ROSE FULBRIGHT (May 2026), https://perma.cc/2GUY-T4TG; Annette Tyman, Joseph R. Vele & Christopher J. Laudenbach, *Colorado Enacts Artificial Intelligence Replacement Law*, SEYFARTH SHAW LLP (May 22, 2026), https://perma.cc/6LHD-Q875.

82. *See* Sibo Ma, Alejandro Salinas, Peter Henderson & Julian Nyarko, *Breaking Down Bias: On the Limits of Generalizable Pruning Strategies*, 2025 PROCS. ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 2437, 2437, https://perma.cc/ED98-QUFS (explaining that the core tension of AI liability and regulatory frameworks arises from the development of general purpose models that are deployed in domain-specific contexts).

bias auditing, to ensure statistical differences remain harmless.[83] On the other hand, an algorithm that represents historical data but has a negative impact on a legally protected group may demand remediation, potentially through legal and policy avenues that influence technical development. For example, in Meta's 2022 settlement with the DOJ over discriminatory housing advertisements, the company was required to develop a revised algorithm that mitigated discriminatory advertisement delivery based on race and sex.[84] The revised algorithm mitigated bias by approximating protected characteristics of users in the aggregate to equalize advertisement delivery, rather than directly taking a user's protected characteristics into account.[85]

This has implications for how AI is steered, whether in service of representational goals or anticipation of or response to an algorithm's disparate impact. If current legal doctrine does not allow a developer to consider legally protected characteristics in designing an AI model, then how might they ensure an LDA is deployed? One answer may require technological advancement that is responsive to these legal developments—perhaps by building on the model multiplicity approach, the very kind of groundbreaking work this field needs. Another is thinking more innovatively about the law, whether by developing a new antidiscrimination legal framework or applying a different doctrine altogether. For example, scholars have articulated legal approaches to addressing AI discrimination that would supplement antidiscrimination law, intended to fill in gaps the existing doctrine might not cover. These approaches generally center around holding *developers* rather than just *decisionmakers* accountable for harmful technology, specifically by applying consumer protection, product safety, and products liability law.[86]

Unavoidably, these potential solutions to AI bias are further complicated by a technological landscape that is nearly impossible to keep up with. Traditional definitions of AI bias exist in predictive AI settings, and much of the field had to adjust the quantification and evaluation for biases in the generative AI setting. Now, in a not-too-distant future where AI "agents" evolve from independently performing more mundane tasks to possessing general human intelligence and making their own unprompted high-stakes decisions, it will require an even

83. For an example of how two researchers designed a role-playing scenario to conduct a bias audit of ChatGPT, see Katherine-Marie Robinson & Violet Turri, *Auditing Bias in Large Language Models*, CARNEGIE MELLON UNIV. SOFTWARE ENG'G INST. SEI BLOG (July 22, 2024), https://perma.cc/8Q6U-HEFE.

84. Press Release, U.S. Dep't of Justice, Justice Department and Meta Platforms Inc. Reach Key Agreement as They Implement Groundbreaking Resolution to Address Discriminatory Delivery of Housing Advertisements (Jan. 9, 2023), https://perma.cc/3ZHF-SL4T.

85. GUIDEHOUSE, VRS COMPLIANCE METRICS VERIFICATION 40 (2024), https://perma.cc/6433-BURK.

86. Andrew D. Selbst, *Artificial Intelligence and the Discrimination Injury*, 78 FLA. L. REV. (forthcoming 2026) (manuscript at 56-68) (on file with authors).

more nuanced and careful understanding of each context in which the technology is deployed to unpack and navigate AI bias effectively.[87]

Conclusion

Contemporary disputes over AI bias are best understood as normative disputes over representation—disputes that illuminate the multiplicity of values and contexts in which AI bias can be defined. Systems that reflect the world as it exists may reproduce disparities that are unfair or discriminatory in some contexts, while in others they may be helpful in improving the human condition. Both scenarios raise complex legal and policy questions about whether and when historical patterns should be preserved or altered in service of representational goals or in response to or in anticipation of disparate outcomes. The task, then, is not to treat AI bias as a settled concept—which may limit the types of interventions that are needed in an evolving technological future—but to develop an inquiry-based, context-specific approach to unpacking AI bias. Ultimately, creating a common set of questions to navigate AI bias will help determine when skewed algorithmic outcomes might reflect acceptable tradeoffs or warrant correction and how competing values like accuracy, fairness, and equality should be balanced across different domains.

87. *See* Beth Stackpole, *Agentic AI Explained*, MIT SLOAN SCH. MGMT. (Feb. 18, 2026), https://perma.cc/8PBV-Q8UK; *see also* Addison J. Wu, Ryan Liu, Xuechunzi Bai & Thomas L. Griffiths, *Large Language Models Develop Novel Social Biases Through Adaptive Exploration*, 2026 PROCS. INT'L CONF. ON MACHINE LEARNING 306, 306, https://perma.cc/G85H-J89N (showing that as large language models are put into agents that make decisions, they can develop "novel social biases about artificial demographic groups even when no inherent differences exist").