Harvard Journal of Law & Technology
Volume 38, Number 4 Spring 2025

FORESEEABLY DISCRIMINATORY: MODERNIZING THE EQUAL PROTECTION DOCTRINE FOR AI SYSTEMS

Justin Curl & Emily Hua***

ABSTRACT

The adoption of AI systems threatens the continued viability of an Equal Protection jurisprudence centered on discriminatory intent and disparate impact. Unlike human actors, these systems lack independent intent and may conceal policymakers' biases by using proxies for protected classes. But even as these characteristics complicate existing doctrine, others highlight a path forward.

We argue that, in the context of AI systems, the Supreme Court should incorporate the foreseeability of disparate impact into its discriminatory intent analysis for two reasons. First, disparate impact created by an AI system likely reflects an intentional choice by policymakers. Development practices like predicting model performance with test sets and training multiple model versions allow policymakers to repeatedly adjust a model until its actual outcomes match their intended ones. Second, an alternative method for finding discriminatory intent is necessary to offset the lack of traditional evidence created by an AI system for a reviewing court.

We offer two ways the Court can update its doctrinal analysis and apply both versions to a real-world case study: The Allegheny Family Screening Tool ("AFST"). This system, launched by Allegheny County in 2016, assists child welfare workers in deciding whether an allegation of abuse or neglect requires further investigation. Increasingly, systems like AFST are determining the most important outcomes of our lives. Without changes to the Court's doctrine, these systems can be used to discriminate with impunity, leaving the Fourteenth Amendment's core promise of equality unfulfilled.

---

* Harvard Law School, Candidate for J.D., 2026.

** Harvard Law School, Candidate for J.D., 2025. Thank you to the editors of *JOLT*, particularly Jane Boettcher, for thoughtful edits, to Professor Guy-Uriel Charles for inspiring us to write this paper, to Professor Aileen Nielsen and Professor Martha Minow for helpful feedback and early confidence in our ideas, and to Armin Ameri, Lev Cohen, and Nina Gayleard for their excellent comments. All errors are our own.

TABLE OF CONTENTS

I. INTRODUCTION 1168

II. THE MISFIT BETWEEN ANTI-DISCRIMINATION LAW AND AI SYSTEMS 1172
A. Background on Anti-Discrimination Law 1173
1. Disparate Impact 1173
2. Discriminatory Intent 1174
3. Inferring Discriminatory Intent From Disparate Impact 1175
B. AI Systems Undermine Existing Discriminatory Intent Doctrine 1175

III. OUR PROPOSAL: ANALYZING THE FORESEEABILITY OF DISPARATE IMPACT 1180
A. The Role of Foreseeable Disparate Impact in Finding Discriminatory Intent 1180
1. Defining Foreseeability 1180
2. The Court's Uncertain Stance on the Foreseeability of Disparate Impact 1181
B. Why Courts Should Analyze the Foreseeability of Disparate Impact for AI Systems 1183
1. Policymakers Can Control the Outcomes of AI Systems More Than Human Ones 1183
a. Test Sets Allow Policymakers to Predict Disparate Impact Before Deployment 1183
b. Developers Can Choose Between Multiple, Similarly Performant Options 1185
2. AI Systems Generate a Limited Record for Reviewing Courts 1187
C. How to Incorporate Foreseeability into the Existing Doctrine 1188

IV. ALLEGHENY COUNTY FAMILY SCREENING TOOL CASE STUDY 1190
A. Background on Allegheny Family Screening Tool 1190
B. Applying the Foreseeability Analysis to the Allegheny Family Screening Tool 1193

V. CONCLUSION 1197

I. INTRODUCTION

Antidiscrimination law prohibits decision-makers from treating people differently based on race, religion, sex, or other protected characteristics. Over the past sixty years, a legal framework for identifying impermissible discrimination has evolved around the concepts of

discriminatory intent and disparate impact. Yet the continued viability of these concepts is under threat from the adoption of artificial intelligence ("AI") systems.

AI systems are replacing human decision-makers in important contexts like criminal sentencing,[1] hiring,[2] access to credit,[3] and child welfare.[4] These decision-making algorithms use machine learning to predict future outcomes about an individual,[5] often by learning from a labeled dataset of example inputs and corresponding "correct" outputs.[6]

---

1. *See* Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/3MZ9-47ED] (describing the COMPAS risk assessment system and finding that the software "was particularly likely to falsely flag black defendants as future criminals"); Kashmir Hill, *Facial Recognition Led to Wrongful Arrests. So Detroit Is Making Changes*, N.Y. TIMES (June 29, 2024), https://www.nytimes.com/2024/06/29/technology/detroit-facial-recognition-false-arrests.html [https://perma.cc/Q56L-DYXJ].

2. *See* Aaron Rieke & Miranda Bogen, *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*, UPTURN 14–16 (Dec. 10, 2018), https://www.upturn.org/work/help-wanted/ [https://perma.cc/K5P8-S6TR] (detailing the usage of predictive hiring algorithms used by employers).

3. *See* Laura Blattner & Scott Nelson, How Costly is Noise? Data and Disparities in Consumer Credit 26 (May 5, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/2105.07554 [https://perma.cc/JF3M-6NWC] (discussing the usage of algorithms in the generation of credit scores and the disparities present for disadvantaged groups); Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 157 (2016).

4. *See* Rhema Vaithianathan, Emily Kulick, Diana Benavides Prado & Emily Putnam-Hornstein, *Allegheny Family Screening Tool: Methodology, Version 2*, CTR. FOR SOC. DATA ANALYTICS 2 (2019), https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V2-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-7.pdf [https://perma.cc/6UK7-VKEK] [hereinafter *AFST Version 2*].

5. *See* Angelina Wang, Sayash Kapoor, Solon Barocas & Arvind Narayanan, *Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy*, 1 ACM J. ON RESPONSIBLE COMPUTING 9:1, 1 (2024).

6. The choice of outcome variable is very important. Sometimes there might not be data that directly measures the outcome the developer is trying to predict. For example, in the criminal justice context, a developer might try to predict the likelihood of someone committing a crime but may only have data about arrests, which does not always reflect the true rate of criminal activity. *See* Lisa Stolzenberg, Stewart J. D'Alessio & Jamie L. Flexon, *The Usual Suspects: Prior Criminal Record and the Probability of Arrest*, 24 POLICE Q. 31, 31 (2021). Additionally, poor choice of outcome variable may lead to undesirable feedback loops, in which an algorithm's predictions amplify existing biases. *See, e.g.*, Nathan Matias & Lucas Wright, *Impact Assessment of Human-Algorithm Feedback Loops*, JUST TECH SOC. SCI. RSCH. COUNCIL (Mar. 1, 2022), https://just-tech.ssrc.org/field-reviews/impact-assessment-of-human-algorithm-feedback-loops/ [https://perma.cc/K633-NR95]; Jessica Saunders, Priscillia Hunt & John S. Hollywood, *Predictions Put Into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. EXPERIMENTAL CRIMINOLOGY 347, 347 (2016); David Hickton & Beth Schwanke, *Comment Letter on National Telecommunications and Information Administration Artificial Intelligence Accountability Policy Request for Comment* (June 11, 2023), https://downloads.regulations.gov/NTIA-2023-0005-0636/attachment_1.pdf [https://perma.cc/8Z6L-KRR2]. For more extensive analysis on why the choice of outcome variable is both important and potentially problematic, see Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 812 (2020); Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. L. REV. 649, 665 (2021).

For example, in the hiring context, the inputs might be a candidate's previous employment history, while the outcome to be predicted might be whether the candidate was hired.[7]

One well-documented algorithmic decision system[8] is the Allegheny Family Screening Tool ("AFST"), which was launched by Allegheny County ("the County") in 2016 to assist its child welfare workers.[9] AFST leverages a wide range of existing data on reported children, such as behavioral health and juvenile probation records, to generate risk scores that child welfare workers use in deciding whether to recommend a reported child's family for further screening.[10] Though lauded for efficiency improvements,[11] AFST has been criticized for its use of proxy variables that may reinforce racial or socioeconomic disparities[12] and was under investigation by the Justice Department for potential discrimination against families with disabilities.[13]

AFST is part of a wider push by U.S. government agencies to deploy algorithmic decision systems — local or state child welfare agencies in at least twenty-six states and D.C. are considering such tools, and jurisdictions in at least eleven of these states have already adopted

---

7. Rieke & Bogen, *supra* note 2.

8. Algorithmic decision systems refer to the subset of AI systems that make decisions about people based on predictions made by algorithmic techniques, such as supervised machine learning. Although generative AI systems, particularly large language models like ChatGPT, can be used to make decisions about people, these are not the type of AI systems that are generally referred to as algorithmic decisions systems and are outside the scope of this paper. *See* David Restrepo Amariles, *Algorithmic Decisions Systems: Automation and Machine Learning in the Public Administration*, in THE CAMBRIDGE HANDBOOK OF THE LAW OF ALGORITHMS 273, 277 (Woodrow Bartfield ed., 2020). Also, because "algorithmic decision system" is a clunky phrase, we use the terms AI system, AI decision-making system, and algorithmic decision system interchangeably.

9. *AFST Version 2*, *supra* note 4, at 2.

10. *Id.* at 3.

11. Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko & Rhema Vaithianathan, *A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions*, 81 PROC. MACH. LEARNING RSCH. 134, 136 (2018) (finding that though data was available before, it was "challenging for County staff to efficiently access, review, and make meaning of all available records").

12. *See* Emnet Almedom, Nandita Sampath & Joanne Ma, *Algorithms and Child Welfare: The Disparate Impact of Family Surveillance in Risk Assessment Technologies*, BERKELEY PUB. POL'Y J., Fall 2020, at 22–23; Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1876, 1893 (2020); Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov et al., *The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool*, 2023 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1292, 1298–99; *see, e.g.*, VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 155 (2018).

13. *See* Sally Ho & Garance Burke, *Here's How an AI Tool May Flag Parents with Disabilities*, AP NEWS (Mar. 15, 2023), https://apnews.com/article/child-protective-services-algorithms-artificial-intelligence-disability-02469a9ad3ed3e9a31ddae68838bc76e [https://perma.cc/TD6C-VLGV].

them.[14] As long as these AI systems are used, courts will need to resolve an increasing number of challenges under the Fourteenth Amendment's Equal Protection Clause. Unfortunately, the Supreme Court's current jurisprudence centered around discriminatory intent is ill-suited for algorithmic contexts.[15] These algorithmic decision-making systems differ from their human counterparts in two key ways: First, they do not possess intent of their own, forcing courts to rely on the intent of their human creators.[16] Second, they can better identify correlations between protected (e.g., race, religion, sex, etc.) and unprotected characteristics (e.g., height, educational background, gambling history, etc.), which enables developers to hide discriminatory intent through proxies for sensitive characteristics.[17] But even as these characteristics complicate the existing doctrine, others highlight a path forward.

In this paper, we argue that, in the specific context of AI systems, courts should incorporate the foreseeability of disparate impact into their discriminatory intent analysis. The Court contemplated this question in the Equal Protection context in *Personnel Administrator of Massachusetts v. Feeney*.[18] It rejected inferring discriminatory intent solely from a decision-maker's awareness that a specific practice might produce a disparate impact.[19] But *Feeney* was a fact-specific ruling and may not apply to cases where the disparate impact is so clear that it does not require any inferences by the policymaker. And even if *Feeney* did set down a general rule against inferring discriminatory intent from foreseeable disparate impact, we offer two justifications for creating a narrow carveout for AI systems.

First, policymakers have greater control over policy outcomes when they use AI decision-making systems. The common industry practices of predicting performance with test sets and training multiple model types allow policymakers to repeatedly adjust a model until its predicted outcomes match their intended ones.[20] Because policymakers have greater control over the consequences of a policy when they use AI systems, it is more fair to treat those consequences as intended by the policymaker.[21]

Second, there is a diminished record for reviewing courts to consider when analyzing for discriminatory intent. Unlike human decision-making systems where there are two layers of records (one when the

14. *See* Anjana Samant, Aaron Horowitz, Kath Xu & Sophie Beiers, *Family Surveillance by Algorithm*, AM. C.L. UNION 3 (Sept. 29, 2021), https://www.aclu.org/documents/family-surveillance-algorithm [https://perma.cc/27QT-WJBF].

15. *See infra* Section II.B.

16. *Id.*

17. *Id.*

18. 442 U.S. 256 (1979).

19. *Id.* at 279.

20. *See infra* Section III.B.1.b.

21. *See infra* Section III.B.2.

policymakers write the policy and another when the implementing decision-makers explain their reasoning), AI decision-making systems only generate one layer of records: whatever the policymakers and developers disclose when designing the model.[22] By collapsing the writing and implementation stages of decision-making practices, AI systems reduce the record available for reviewing courts, so it makes sense to correct for the reduced availability of evidence by lowering the bar for what needs to be found (i.e., by making it easier to find discriminatory intent).[23]

Courts can incorporate foreseeability into its doctrinal analysis in two ways. First, courts could treat foreseeable disparate impact during deployment of an AI model as sufficient evidence for finding discriminatory intent. This approach would be more disruptive because it conflicts with some interpretations of *Feeney*. Second, courts could include foreseeable disparate impact as a source of evidence in the existing test for inferring discriminatory intent from circumstantial evidence articulated in *Village of Arlington Heights v. Metropolitan Housing Development Corporation*.[24] The approach would require far less doctrinal change.

The remainder of the Note will proceed as follows: Part II describes the paper's motivation by explaining how AI systems challenge existing tests for discriminatory intent. Part III introduces our proposed solution of analyzing the foreseeability of disparate impact in the Court's test for discriminatory intent, explains why this solution makes sense, and offers two ways our proposal can be incorporated into Equal Protection doctrine. Part IV introduces AFST as a case study and applies both doctrinal versions of our proposal.

II. THE MISFIT BETWEEN ANTI-DISCRIMINATION LAW AND AI SYSTEMS

AI systems challenge existing anti-discrimination law. Section II.A introduces the current state of anti-discrimination law, with a focus on the Equal Protection Clause, while Section II.B explains how AI systems render the existing test for discriminatory intent inapposite by allowing developers to hide intent with proxies for sensitive characteristics.

---

22. *See infra* Section III.B.2.

23. *Id.*

24. 429 U.S. 252, 267–68 (1977) (listing (1) "[t]he historical background of the decision," (2) "[t]he specific sequence of events leading up to the challenged decision," (3) "[d]epartures from normal procedural sequences," and (4) "contemporary statements by members of the decision-making body, minutes of its meetings, or reports" as evidence for analysis).

A. Background on Anti-Discrimination Law

Current anti-discrimination law is rooted in federal legislation and the Constitution. Statutorily, three key pieces of legislation dating back to the civil rights era — Title VII of the Civil Rights Act of 1964,[25] the Voting Rights Act of 1965,[26] and the Fair Housing Act of 1968[27] — collectively prohibit treating people differently on the basis of protected characteristics like race and religion.[28] Constitutionally, under the Equal Protection Clause of the Fourteenth Amendment, the Court will invalidate state action[29] that is facially neutral only when it results in racially disparate impact[30] and was motivated by discriminatory intent.[31] This Note focuses on the Court's current test for discriminatory intent under the Equal Protection Clause. Thus, while our analysis is relevant to all anti-discrimination law, it is primarily applicable to the decisions of governments and government contractors — decisions which qualify as state action and are thus subject to the Fourteenth Amendment.

1. Disparate Impact

A law or policy produces a disparate impact when it "bears more heavily on one race than another."[32] In making this judgment, courts commonly compare the demographics of the population receiving a benefit with the broader region's population demographics (e.g., comparing the percentage of students of a particular race admitted to a prestigious high school with the percentage of people of that race in that

25. 42 U.S.C. §§ 2000e–2000e17 (2018).

26. 52 U.S.C. § 10101 (2018).

27. 42 U.S.C. § 3601–3619 (2018).

28. *See* Eric K. Yamamoto, *Critical Race Praxis: Race Theory and Political Lawyering Practice in Post-Civil Rights America*, 95 MICH. L. REV. 821, 844–45 (1997).

29. The Equal Protection Clause only applies to state action. Virginia v. Rives, 100 U.S. 313, 318 (1879) ("The provisions of the Fourteenth Amendment of the Constitution we have quoted all have reference to State action exclusively, and not to any action of private individuals.").

30. *See* Palmer v. Thompson, 403 U.S. 217, 224 (1971) ("But no case in this Court has held that a legislative act may violate equal protection solely because of the motivations of the men who voted for it.").

31. Washington v. Davis, 426 U.S. 229, 239 (1976) ("But our cases have not embraced the proposition that a law or other official act, without regard to whether it reflects a racially discriminatory purpose, is unconstitutional *solely* because it has a racially disproportionate impact."); Vill. of Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252, 264–65 (1977) ("[O]fficial action will not be held unconstitutional solely because it results in a racially disproportionate impact."); *see also* Stephen M. Rich, *Inferred Classifications*, 99 VA. L. REV. 1525, 1525–26 (2013) ("[F]acially neutral legislation that produces racially disproportionate harms will not violate equal protection unless motivated by a discriminatory purpose.").

32. *Washington v. Davis*, 426 U.S. at 242; *see also* Coal. for TJ v. Fairfax Cnty. Sch. Bd., 218 L. Ed. 2d 71, 74 (2024) (Alito, J., dissenting from denial of certiorari).

school district).[33] Alternatively, courts will sometimes evaluate disparate impact by comparing the demographics of the population receiving a benefit under the current policy with the demographics under the old policy.[34] However, some have criticized this before-and-after comparison for turning "the previous status quo into an immutable quota."[35]

2. Discriminatory Intent

The Court's articulation of discriminatory intent "implies more than intent as volition or intent as awareness of consequences;" it requires that policymakers made a decision "at least in part 'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group."[36] Importantly, the Court "does not require a plaintiff to prove that the challenged action rested solely on racially discriminatory purposes,"[37] just "that a discriminatory purpose has been a motivating factor in the decision."[38] Though discriminatory intent may be satisfied by direct evidence, it "need not be proved by direct evidence."[39] "Necessarily, an invidious discriminatory purpose may often be inferred from the totality of the relevant facts."[40] The *Arlington Heights* Court listed the following as sources of evidence to infer discriminatory purpose: (1) "[t]he historical background of the decision," (2) "[t]he specific sequence of events leading up to the challenged decision," (3) "[d]epartures from normal procedural sequences," and (4) "contemporary statements by members of the decision-making body, minutes of its meetings, or reports."[41]

---

33. *See, e.g.*, Arce v. Douglas, 793 F.3d 968, 978 (9th Cir. 2015) (finding disparate impact where an Arizona statute prohibited ethnic studies programs, where ninety percent of participating students were of Mexican or other Hispanic descent); Coal. for TJ v. Fairfax Cnty. Sch. Bd., 68 F.4th 864, 879 (4th Cir. 2023), *cert. denied*, 218 L. Ed. 2d 71 (2024) (finding no disparate impact of a school policy on White and Asian-American students where "those students have had greater success in securing admission to [the school] under the policy than students from any other racial or ethnic group" despite a drop in percentage of both groups after the implementation of the policy).

34. Coal. for TJ v. Fairfax Cnty. Sch. Bd., 2022 WL 579809, at *6 (E.D. Va. Feb. 25, 2022), *rev'd and remanded*, 68 F.4th 864 (4th Cir. 2023) ("The proper method for determining the 'impact of the official action' is a simple before-and-after comparison.") (quoting *Arlington Heights*, 429 U.S. at 266).

35. *See TJ*, 68 F.4th at 881.

36. Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256, 279 (1979).

37. *Arlington Heights*, 429 U.S. at 265.

38. *Id.* at 265–66.

39. Rogers v. Lodge, 458 U.S. 613, 618 (1982).

40. Washington v. Davis, 426 U.S. 229, 242 (1976).

41. *Arlington Heights*, 429 U.S. at 267–68. *See, e.g.*, Flores v. Pierce, 617 F.2d 1386, 1389 (9th Cir. 1980) (applying *Arlington Heights* factors to "infer an intent to disguise a racial animus").

3. Inferring Discriminatory Intent From Disparate Impact

In cases without evidence of explicit discriminatory intent, the Court can still infer discriminatory intent when "a clear pattern, unexplainable on grounds other than race, emerges from the effect of the state action . . . . But such cases are rare."[42] The Court only infers discriminatory intent from disparate impact in egregious cases, such as *Yick Wo v. Hopkins*[43] and *Gomillion v. Lightfoot*.[44] In the former, the Court invalidated a policy prohibiting laundries from operating out of wooden buildings.[45] Though the policy was facially neutral, only one permit application of the 200 Chinese owned laundries was granted; only one non-Chinese owned laundry permit application was denied.[46] And in the latter, a policy change modified a voting district from "a square . . . into a strangely irregular twenty-eight-sided figure" and effectively removed "all save four or five of its 400 [Black] voters while not removing a single white voter or resident."[47] "Absent a pattern as stark as that in *Gomillion* or *Yick Wo*, impact alone is not determinative."[48]

B. AI Systems Undermine Existing Discriminatory Intent Doctrine

AI systems complicate existing Equal Protection doctrine.[49] Two differences between algorithmic decision systems and human ones explain this complication: (1) algorithms do not possess intent of their own,[50] and (2) developers can hide discriminatory intent with proxies.[51] In this Section, we briefly explain how an AI system's lack of intent

42. *Arlington Heights*, 429 U.S. at 266; *see also* Michael Selmi, *Proving Intentional Discrimination: The Reality of Supreme Court Rhetoric*, 86 GEO. L.J. 279, 304 (1997).

43. 118 U.S. 356 (1886).

44. 364 U.S. 339, 341 (1960).

45. *Yick Wo*, 118 U.S. at 374.

46. *Id.* at 359.

47. *Gomillion*, 364 U.S. at 341.

48. Vill. of Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252, 266 (1977).

49. *See, e.g.*, Note, *Beyond Intent: Establishing Discriminatory Purpose in Algorithmic Risk Assessment*, 134 HARV. L. REV. 1760, 1766–69 (2021); Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 311–21 (2020) (explaining that excluding protected characteristics from algorithmic decision systems does not eliminate racial disparities and the difficulty of dealing with proxy effects); Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 116 (2018) (noting that "people themselves may not know why and how they are choosing"); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1083 (2019); Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 890, 892–93 (2018) (describing how legal doctrines apply inadequately to AI due to the black box nature of its decision-making process).

50. *Beyond Intent*, *supra* note 49, at 1766; *id.* at 1766 n.42 (citing Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1066 (2019)).

51. Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1265 (2020).

makes it harder to satisfy the Equal Protection discriminatory intent requirement before describing how AI developers can use AI models to circumvent the Court's test for discriminatory intent. Then, because much prior academic work on proxies was published before the Court's decisions in *Students for Fair Admissions, Inc. v. Presidents & Fellows of Harvard College*[52] ("*SFFA*") and *Coalition for TJ v. Fairfax County School Board*[53] ("*TJ*"), we use these recent cases to explain why the Court struggles to address proxies coherently.

For human decision-makers, courts can rely on direct and circumstantial evidence of discriminatory intent. While there is always the risk that evidence of intent is contrived or stated reasons are pretextual, courts have a long history of thinking about when and how to ascribe intent to human actors. AI systems undermine this practice because they lack intent.[54] Further, even if they had intentions,[55] there has not been much work on techniques for ascertaining a model's "intent,"[56] and courts have little-to-no experience evaluating the technical details of AI models. Because of this difficulty, courts are forced to rely solely on the intent of the policymakers and developers who deploy these systems.

But AI systems make it easier for policymakers and developers to hide discriminatory intent. These systems can better detect correlations between variables than human decision-makers can and can thus identify and use proxies for sensitive characteristics.[57] Even if an algorithm

52. 600 U.S. 181 (2023).

53. 218 L. Ed. 2d 71 (2024).

54. Ian Ayres & Jack M. Balkin, *The Law of AI is the Law of Risky Agents Without Intentions*, U. CHI. L. REV. ONLINE 1 (Nov. 2024) (discussing the difficulty of ascribing intent to AI programs, as "the law of AI is the law of risky agents without intentions").

55. As far as the authors are aware, no one suggests that predictive AI systems have intent. That said, we recognize that people disagree about whether large language models can be said to have intent, though this issue is outside the scope of this paper. *Compare* Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 610, 616 ("Text generated by a[] [language model] is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind.") *with* Rebecca L. Perez, *Do Language Models Communicate? Communicative Intent and Reference from a Derridean Perspective*, 2 Технологии в инфосфере [TECH. & LANGUAGE] 40, 40 (2024) ("Based on these concepts, the paper argues that [language models] could be said to communicate and understand.").

56. *Cf.* Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen et al., *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*, ANTHROPIC (May 21, 2024), https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html [https://perma.cc/2UCD-S87T] (explaining that current research does not adequately address limitations to mechanistic interpretability).

57. Prince & Schwarcz, *supra* note 51, at 1263; *see* Aaron S. Coyner, Praveer Singh, James M. Brown, Susan Ostmo, RV Paul Chan, Michael F. Chiang et al., *Not Color Blind: AI Predicts Racial Identity from Black and White Retinal Vessel Segmentations* 2 (2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2109.13845 [https://perma.cc/PDZ9-WTM7] (finding that a visual recognition model could reliably infer race from retinal images).

cannot directly access a sensitive characteristic as an input variable, its predictions may still indirectly rely on that characteristic.[58] For example, with AFST, one of the input variables is the mother's level of education.[59] The variable is "1" if she received an "Associate degree," "Bachelor's degree," "Master's degree" or "Doctorate or professional degree," and "0" otherwise.[60] Since the distribution of educational attainment in the United States varies based on race,[61] a malicious developer could hide their discriminatory intent by constructing a model that produces disparate impact but makes decisions based on a facially-neutral variable like education level.[62] But oftentimes, developers do not actively choose which variables to include. Instead, they rely on feature selection algorithms to identify the set of variables whose inclusion maximizes accuracy.[63] AFST's developers, for example, relied on a feature selection algorithm to narrow 800 candidate variables to a final subset of seventy-two variables.[64] In such cases, developers may unintentionally discriminate based on race. Either way, as long as a developer building an algorithmic decision system does not leave a record of discriminatory intent, the Court's current test will conclude that there is no discriminatory intent, regardless of the developer's true intentions.

During oral arguments for *SFFA*, Justice Kavanaugh expressed concern about the uncertain status of proxy variables in the Court's affirmative action jurisprudence.[65] He repeatedly asked if it would be constitutional for a school to consider someone's status as a "descendant of slaves" during the admissions process.[66] Each time, the advocates for Students for Fair Admissions, Inc. expressed doubt that affirmative action based on someone's status as a descendant of slaves would be constitutional because the status is so heavily correlated with race.[67] But despite this uncertainty, the Court's ruling ultimately side-stepped

58. Yang & Dobbie, *supra* note 49, at 335–42.

59. *AFST Version 2*, *supra* note 4, at 19.

60. *Id.*

61. *See* Press Release, United States Census Bureau, Census Bureau Releases New Educational Attainment Data (Feb. 24, 2022), https://www.census.gov/newsroom/press-releases/2022/educational-attainmenxt.html [https://perma.cc/YKH9-M6FK].

62. Michael Brenner, Jeannie S. Gersen, Michael Haley, Matthew Lin, Amil Merchant, Richard J. Millett et al., *Constitutional Dimensions of Predictive Algorithms in Criminal Justice*, 55 HARV. C.R.-C.L. L. REV. 267, 294–96 (2020).

63. *See Feature Selection*, HEAVY.AI, https://www.heavy.ai/technical-glossary/feature-selection [https://perma.cc/5Y3P-QXU8].

64. Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand & Tim Maloney, *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation*, CTR. FOR SOC. DATA ANALYTICS 14 (2017), https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V1-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL.pdf [https://perma.cc/66GC-6FUU] [hereinafter *AFST Version 1*].

65. Transcript of Oral Argument at 44–46, Students for Fair Admissions, Inc. v. President of Harv. Coll., 600 U.S. 181 (2023) (No. 21-707).

66. *Id.* at 44.

67. *Id.* at 44–45.

the question of proxy variables highly correlated with a suspect class. The only guidance the Court offered was that a "student must be treated based on his or her experiences as an individual — not on the basis of race."[68] By denying certiorari in *TJ*, the Court again declined to address proxy variables.[69] This silence is especially problematic for algorithmic decisions systems, which can use proxies for protected classes that human beings would otherwise be unable to detect.[70]

Going forward, to address the question of proxies, the Court must either create a bright-line rule allowing all proxies, no proxies, or some proxies, or address them on a case-by-case basis.[71] Each of these options has glaring flaws, which may explain why the Court has struggled to resolve this doctrinal issue.

Treating *all* proxy variables as acceptable, no matter how correlated they are with a protected feature, violates common sense. If courts prevent a state actor from making a decision on the basis of race, it would be logically inconsistent to allow that actor to make that decision on some other basis that is 100% correlated with race. While such a decision would nominally be race-neutral by satisfying "racial blindness," it would violate the principle of individualized treatment because it would treat people not as individuals but as members of the proxy class, which maps perfectly onto race.

Yet treating *all* proxies as *un*acceptable is equally undesirable. Requiring developers to remove all variables that are correlated with a protected characteristic like race will make algorithms far less accurate, since almost every variable in American society is correlated with race.[72] A more targeted solution focused on close proxies is similarly

68. Students for Fair Admissions, Inc. v. President of Harv. Coll., 600 U.S. 181, 231 (2023).

69. Coal. for TJ v. Fairfax Cnty. Sch. Bd., 218 L. Ed. 2d 71 (2024).

70. Prince & Schwarcz, *supra* note 51, at 1265–66.

71. Scholars disagree over whether the incorporation of race in algorithms at various stages is actually per se unconstitutional. Some believe that the usage of race at any point poses strong constitutional problems. *See* Daniel Ho & Alice Xiang, *Affirmative Algorithms: The Legal Grounds for Fairness as Awareness*, U. CHI. L. REV. ONLINE 134, 138–40 (Oct. 2020) (analyzing how common "group fairness" approaches to equality that take into account race pose Equal Protection problems); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 819 (2014) ("There is a strong case that most or all of the risk prediction instruments now in use are unconstitutional . . . ."). Others offer potential workarounds to either "zero-out" the effect of race or limit the effect of race in proxy variables so that there is no constitutional issue. *See* Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action*, 110 CAL. L. REV. 1539, 1575–78 (2022) (suggesting various de-biasing strategies); Yang & Dobbie, *supra* note 49, at 346–50 (proposing two algorithmic solutions and analyzing them in the context of the Equal Protection Clause).

72. Yang & Dobbie, *supra* note 49, at 311–19 (discussing various commonly used variables and ways in which they could be a proxy for race); *see also* Prince & Schwarcz, *supra* note 51, at 1306–07 ("[T]his approach would only allow actors to use pre-approved variables. It would thus limit algorithmic proxy discrimination by making AI almost completely useless relative to traditional statistical methods given the availability of so few variables").

impracticable because "prohibiting the use of protected class variables and close proxies does little to prevent intentional obfuscation of biased algorithms" by ill-intentioned developers.[73] A decrease in an algorithmic decision system's overall accuracy, with little countervailing improvement in fairness, would also be practically harmful because it would (1) inefficiently allocate resources and (2) harm public trust in government, as policymakers would be prevented from implementing more effective and efficient algorithms.[74]

Thus, the Court's most realistic option would be to allow proxies some of the time. But distinguishing between acceptable and unacceptable proxies in a principled and consistent manner remains a difficult and open question.[75] For example, when does a proxy or a collection of proxies for race become an important enough factor in a decision for that decision to be made "on the basis of race?"[76] Is an algorithmic decision made on the basis of race when proxies are given sole weight, majority weight, or more weight than any other variable? And does the degree of correlation between a proxy variable and sensitive characteristic matter for the analysis? For example, should a proxy that is 99% correlated with race be treated differently than one that is 51% correlated with race?

Alternatively, courts might prohibit developers from using proxies on a case-by-case basis. Courts could independently determine whether an input variable, such as a mother's education level, is an acceptable basis for an algorithm's prediction. Professor Sandra Wachter proposes some principles for evaluating input variables: "(1) stability, (2) transparency, (3) empirical coherence, and (4) normative and ethical acceptability."[77] While such principles may guide court action, this approach ultimately requires courts to make difficult policy judgments that may be better left to the legislature. Moreover, because models can have hundreds of input variables, each potentially requiring individualized

---

73. Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. L. REV. 649, 669–670 (2021); *see id.* at 670 (describing the obfuscation techniques to "conceal whether an algorithmic decision-making process is driven by protected class variables").

74. *See* Gabriel Scheffler & Daniel E. Walters, *The Submerged Administrative State*, 2024 WIS. L. REV. 789, 792–93 (2024) (explaining that a belief that the government "simply 'does not work'" leads to a loss of "trust in government").

75. *See* Prince & Schwarcz, *supra* note 51, at 1306–17 (highlighting the difficulties of five proposed solutions to the problem of proxy discrimination).

76. Students for Fair Admissions, Inc. v. President of Harv. Coll., 600 U.S. 181, 231 (2023) ("In other words, the student must be treated based on his or her experiences as an individual — not on the basis of race.").

77. Sandra Wachter, *The Theory of Algorithmic Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law*, 97 TUL. L. REV. 149, 198 (2022).

consideration, this approach risks creating an unworkable standard that would overwhelm judges with litigation.[78]

Overall, none of these approaches to proxies are satisfying. As a result, algorithmic decision systems, which allow government actors to hide discriminatory intent using proxies, function as a glaring gap in the Court's current intent-focused anti-discrimination analysis. Fortunately, there are other aspects of algorithmic decision systems, which we discuss in the next Part, that provide a path forward for the Court's anti-discrimination jurisprudence.

III. OUR PROPOSAL: ANALYZING THE FORESEEABILITY OF DISPARATE IMPACT

Part III introduces our solution to the problems described in Part II. We recommend that courts consider foreseeable disparate impact in their analyses of discriminatory intent for AI systems. We begin in Section III.A by examining the Court's current treatment of foreseeable disparate impact. Then, in Section III.B, we explain the unique aspects of AI systems that justify an inference of discriminatory intent from foreseeable disparate impact. Finally, in Section III.C, we offer two ways — one more and one less doctrinally disruptive — for courts to incorporate the foreseeability of disparate impact into their Equal Protection analysis.

A. The Role of Foreseeable Disparate Impact in Finding Discriminatory Intent

1. Defining Foreseeability

The Court defined foreseeability in *Personnel Administrator of Massachusetts v. Feeney* as an "awareness of consequences."[79] In *Feeney*, the Court evaluated whether a Massachusetts statute that strictly advantaged veterans over nonveterans in civil service appointments violated the Equal Protection Clause.[80] The trial court "found that the absolute preference . . . [had] a devastating impact on the employment opportunities for women"[81] because of the large veteran population and the fact that most veterans were men.[82] The Court

78. *Cf.* Andrew Coan & Nicholas Bullar, *Judicial Capacity and Executive Power*, 102 VA. L. REV. 766, 769 (2016) (explaining how the judicial capacity model suggests that the Court is less deferential in cases with "clear, categorical" rules, which decreases litigation).

79. Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256, 279 (1979).

80. *Id.* at 259.

81. *Id.* at 260.

82. *See id.* at 270 (finding that at the time of the litigation "over 98% of the veterans in Massachusetts were male; only 1.8% were female" and "over one-quarter of the Massachusetts population were veterans").

concluded that legislative decisions could foreseeably create disparate impact because the choice to absolutely preference veterans in hiring led to "an adverse impact upon nonveterans . . . . And it cannot seriously be argued that the Legislature of Massachusetts could have been unaware that most veterans are men."[83] The Court expanded on the definition of foreseeability in *Columbus Board of Education v. Penick*,[84] a case deciding whether a city's conduct had the "purpose and effect of causing and perpetuating segregation in public schools, contrary to the Fourteenth Amendment."[85] The Court adopted the trial court's finding that "the selection of sites for new school construction . . . had the foreseeable and anticipated effect of maintaining the racial separation of the schools."[86] In particular, it highlighted the District Court's finding that "of the 103 schools built by the Board between 1950 and 1975, 87 opened with racially identifiable student bodies and 71 remained that way at the time of trial" and conclusion that "[t]his result was reasonably foreseeable under the circumstances in light of the sites selected, and the Board was often specifically warned that it was, without apparent justification, choosing sites that would maintain or further segregation."[87] The Court also agreed with the Sixth Circuit's conclusion that the "record actually require[d] no reliance upon inference" because it "contain[ed] repeated instances where the Columbus Board was warned of the segregative effect of proposed site choices, and was urged to consider alternatives which could have had an integrative effect."[88] In effect, in cases where a future disparate impact is expressly communicated to decision-makers (*Penick*) or the correlation between a proxy variable and a protected characteristic is obvious (*Feeney*), the Court has concluded the resulting disparate impact is foreseeable.

2. The Court's Uncertain Stance on the Foreseeability of Disparate Impact

Having introduced the Court's definition of foreseeability, we next turn to its treatment of foreseeable disparate impact in its Equal Protection analysis. The Court set the rule in *Feeney*, rejecting the theory of some lower courts that foreseeable disparate impact alone could be

83. *Id.* at 278; *see also id.* ("It would thus be disingenuous to say that the adverse consequences of this legislation for women were unintended, in the sense that they were not volitional or in the sense that they were not foreseeable.").

84. 443 U.S. 449 (1979).

85. *Id.* at 452.

86. *Id.* at 462.

87. *Id.* at 462 n.11 (citing Penick v. Columbus Bd. of Educ., 429 F.Supp. 299, 241–43 (S.D. Ohio 1977)).

88. *Id.* at 462–63 n.11 (citing Penick v. Columbus Bd. of Educ., 583 F.2d 787, 804 (6th Cir. 1977)).

enough to show discriminatory intent.[89] While the Court acknowledged that other areas of law treated "the natural and foreseeable consequences of [a person's] voluntary actions" as intentional,[90] it declined to extend this line of reasoning to the Equal Protection Clause. Instead, it concluded that a requirement of "[d]iscriminatory purpose . . . implies more than intent as volition or intent as awareness of consequences."[91] In other words, a decision-maker acts with discriminatory intent only when she chooses a course of action "at least in part 'because of,' not merely 'in spite of'" its disparate effects on a group.[92]

But the language of *Feeney* suggests it is a narrow and fact-specific holding. Although the Court concluded the disparate "impact [wa]s essentially an unavoidable consequence of a legislative policy that has in itself always been deemed to be legitimate,"[93] it heavily relied on the trial court's factual determination that there was no evidence the legislature intended to discriminate against women and that the legislative purpose was only to improve opportunities for veterans.[94] The Court even acknowledged that "the inevitability or foreseeability of consequences of a neutral rule" can have "bearing upon the existence of discriminatory intent," noting that "when the adverse consequences of a law upon an identifiable group are as inevitable as the gender-based consequences of [this provision], a strong inference that the adverse effects were desired can reasonably be drawn."[95] In doing so, it seemed to preserve the possibility that foreseeable disparate impact could indicate discriminatory intent. But even if *Feeney* is interpreted as a complete rejection of analyzing the foreseeability of disparate impact, the

89. *See, e.g.*, United States v. Bd. of Sch. Comm'rs of Indianapolis, 573 F.2d 400, 413 (7th Cir. 1978) (finding that the government "ignor[ing] less segregative options which would have furthered its policies as effectively as the more segregative option it chose" establishes a "prima facie case of discriminatory intent or purpose," which was "justifiable because governmental institutions must be presumed to have knowledge of the natural and foreseeable consequences of their action or inaction, and because there are rarely significant nonracial reasons for preferring a more rather than less segregative alternative"); United States v. Tex. Educ. Agency, 579 F.2d 910, 913 (5th Cir. 1978) (rejecting the notion that *Washington v. Davis* or *Arlington Heights* "abrogated the principle that an actor is held to intend the reasonably foreseeable results of his actions" since the principle was so "fundamental" that "it would be out of character for the Supreme Court to have disapproved its use in discrimination cases — without explicitly saying so").

90. Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256, 278 (1979); *see* Note, *Proving Discriminatory Intent from a Facially Neutral Decision with a Disproportionate Impact*, 36 WASH. & LEE L. REV. 109, 112, 118–19 (1979).

91. *Feeney*, 442 U.S. at 279.

92. *Id.*

93. *Id.* at 279 n.25.

94. *Id.* at 277 ("The District Court's conclusion that the absolute veterans' preference was not originally enacted or subsequently reaffirmed for the purpose of giving an advantage to males as such necessarily compels the conclusion that the State is [sic] intended nothing more than to prefer 'veterans.'").

95. *Id.* at 279 n.25.

next Section offers two justifications for creating a carveout for AI systems.

B. Why Courts Should Analyze the Foreseeability of Disparate Impact for AI Systems

1. Policymakers Can Control the Outcomes of AI Systems More Than Human Ones

Two technical aspects of AI development make it easier for policymakers to control the outcomes of a decision-making process. First, developers frequently use test sets to estimate an AI model's performance before deployment.[96] This ability to precisely measure and predict the consequences of adopting a policy — here, in the form of an AI model — before that model's deployment allows policymakers to repeatedly adjust a model such that its predicted outcomes more closely match their intended outcomes.[97] Second, developers generally train multiple types of models, giving policymakers the choice between models with similar overall accuracy but different per-group performance.[98] Thus, the decision to select a model with similar overall performance but large between-group disparities is more likely to reflect an active choice by a policymaker to discriminate than the decision to write a policy from scratch.

a. Test Sets Allow Policymakers to Predict Disparate Impact Before Deployment

Predictive algorithms make disparate impact foreseeable because their performance can be efficiently measured with test sets.[99] To train supervised machine learning models (which comprise the majority of federal uses of algorithms),[100] a developer must define a ground-truth

96. Tammy Jiang, Jaimie L. Gradus & Anthony J. Rosellini, *Supervised Machine Learning: A Brief Primer*, 51 BEHAV. THERAPY 675, 682 (2020) ("After data processing and reduction, supervised machine learning methods may be implemented. Ideally, prediction algorithms are developed (i.e., fitted) in one sample and then evaluated (i.e., applied) in an independent sample. In the machine learning literature, this is referred to as having a *training set* versus *test set* of observations." (emphases in original)).

97. *See infra* Section III.B.1.a.

98. *See* Emily Black, Manish Raghavan & Solon Barocas, *Model Multiplicity: Opportunities, Concerns, and Solutions*, 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 850, 853 [hereinafter *Model Multiplicity*].

99. *See, e.g.*, Jiang, Gradus & Rosellini, *supra* note 96, at 682.

100. DAVID F. ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 19 fig.5 (2020).

variable that represents the "correct" optimization.[101] This forces developers to choose between specific evaluation metrics, such as accuracy and false positive rate,[102] which embody different notions of fairness, and weigh differing considerations in their active choice of which variable to optimize.[103] Using these metrics, developers can use test sets of unseen real-world candidates to determine whether the model has overfit its training data or has worse performance for certain demographic groups. As a result, any disparate impact that persists from the testing stage to the deployment stage is more likely to reflect an active choice by the policymaker compared to a human decision-making system. Although human decision-makers technically could estimate their own error rates (e.g., a child welfare screening team could compare its historical predictions with rates of removal from households),[104] such evaluation is easier in the algorithmic context because of a model's ability to quickly and cheaply make a large number of predictions that can be compared with the "correct" outcome.[105] In contrast, asking human decision-makers to make "test" predictions purely to evaluate their own performance costs time and resources. It may be both expensive and wasteful to require human decision-makers to

101. For example, if high risk was defined as a screener opting for further investigation, this could create an undesirable feedback loop since whether a call screener opts for further investigation itself depends on the algorithm's prediction of a child as high risk. In a system where call screeners opt for further investigation of Black children more often, an algorithm will rate Black children as higher risk (they are more likely to be investigated). Call screeners, seeing that Black children are rated as higher risk, will in turn opt for further investigation more frequently. This creates a vicious feedback loop that amplifies any initial biases. *See* Chouldechova, Benavides-Prado, Fialko & Vaithianathan, *supra* note 11, at 4.

102. A false positive rate is defined by how often a system incorrectly predicts someone as deserving the outcome. *Classification: Accuracy, Recall, Precision, and Related Metrics*, GOOGLE, https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall [https://perma.cc/HYB4-66JY] [hereinafter *Metrics Definitions*].

103. Computer scientists and legal scholars alike group these standards in two general categories: group fairness and individual fairness. Group fairness prioritizes statistical parity across protected groups, while individual fairness prioritizes incorrect misclassification of individuals. While group fairness falls more closely in line with the anti-subordination principle of racial equality, individual fairness aligns more closely with individualized treatment. *See* Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67, 68 (2017).

In addition, such notions of fairness are often directly incompatible with each other and are impossible to simultaneously optimize for. *See* Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, PROC. INNOVATIONS THEORETICAL COMPUT. SCI., 2017, at 43:3. Thus, "[t]hese statistical notions were designed to reflect different normative and legal notions of fairness and so the choice of which statistical concept to use will inevitably involve normative and legal dimensions." MacCarthy, *supra* note 103, at 68.

104. For example, AFST looks at whether a child "experiences a safety issue so significant that they would need to be removed from their home" two years following the referral. *AFST Version 2*, *supra* note 4, at 2.

105. *See, e.g.*, Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES, Jan. 5, 2018, at 1–2 (comparing the COMPAS algorithm with a study conducted requiring 400 participants to predict recidivism rate).

generate predictions on hundreds of real-world candidates for the sole purpose of estimating their error rates, especially when there are no guarantees the information would actually improve their future performance. As a result, human decision-making systems might rely on other, more general evaluation methods for quality. For example, a college admissions office, starting from the assumption that a "good admit" is equally likely to be a member of any race, might roughly evaluate whether its admissions process is fair for different groups by comparing the demographic composition of its incoming class to the nation more broadly.[106] This approach, though perhaps more efficient for the college admissions office, does not allow a particular decision-maker to evaluate their own performance.[107] And it does not give any real-time information to policymakers about whether their actual admissions were "good admits."[108]

And for the subset of predictive models that are deterministic (the same input features produce the same output),[109] outcomes are even more foreseeable. If policymakers or developers are concerned with bias against specific groups, they can measure performance on hypothetical candidates who are identical with respect to every input variable except for a sensitive variable (such as race or religion). Because a deterministic model produces the same outputs for a given input,[110] policymakers can be confident that hypothetical candidates treated differently based on the sensitive characteristic during the testing stage would also be treated differently after deployment.

b. Developers Can Choose Between Multiple, Similarly Performant Options

The AI development process also creates opportunities for policymakers to choose between different models. Model developers will

---

106. *See* Students for Fair Admissions, Inc. v. President of Harv. Coll., 600 U.S. 181, 215–16 (2023).

107. *See, e.g.*, *id.* at 295 (2023) (Gorsuch, J., concurring) (explaining the steps that an admissions committee takes, including considering whether there are "dramatic drop off[s] relative to the prior year").

108. Though there are processes and available data through which some colleges evaluate their admissions, such as the Common Data Set, these can only provide backwards-looking reflection on success. *See*, *e.g.*, *University of California Accountability Report 2024*, UNIV. OF CAL. (2024), https://accountability.universityofcalifornia.edu/2024/welcome.html [https://perma.cc/3GA3-GQ8Z] (discussing undergraduate attainment goals, such as student success); COMMON DATA SET 2024–2025, PRINCETON UNIV., https://ir.princeton.edu/sites/g/files/toruqf2041/files/documents/cds_2425_princeton.pdf [https://perma.cc/KZC7-Q4KR] (listing admissions information and post-matriculation student academic achievement).

109. *See Glossary: Deterministic Algorithm*, NAT'L INST. STANDARDS & TECH.: COMPUT. SEC. RES. CTR., https://csrc.nist.gov/glossary/term/deterministic_algorithm [https://perma.cc/Y24A-7873].

110. *Id.*

often test out multiple model architectures for the same task to see which one is most accurate.[111] Among the multiple options created by developers, policymakers will need to select the model performance they most prefer. One response is that policymakers should simply choose whichever model is most accurate, but even when the chief metric is maximizing accuracy, "maximally accurate models can produce varying degrees of demographic disparities."[112] In other words, for any decision-making system, there often exists multiple models with similar levels of accuracy that produce different levels of disparate impact between groups.[113] Others have called the existence of an alternative, similarly performant model a less discriminatory algorithm ("LDA").[114] And in cases where "increased accuracy is achieved through increased model complexity," there may be even "*more* predictive multiplicity" and thus more LDAs to choose from.[115] Importantly, generating alternative models is also becoming more feasible over time. There are many techniques that can be used to create LDAs,[116] such as "exploration along the modeling pipeline."[117] As a result, policymakers will increasingly be able to choose between multiple candidate AI models that have similar levels of performance but vary in their between-group disparity in performance.

Taken together, these two technical features support creating a doctrinal carveout for AI systems. The Court may be reluctant to rely on foreseeability with human decision-making systems because it may seem unfair to hold policymakers accountable for consequences that, though likely, were unintentional.[118] AI models alleviate the unfairness concerns because the consequences of deploying a system are much more predictable. Because of this foreseeability, any amount of disparity resulting from an algorithm is more likely to reflect an intentional choice by developers or policymakers, especially in cases where an LDA exists and was ignored. Although a particular model may be

---

111. *See Model Multiplicity*, *supra* note 98, at 850.

112. *Id.*

113. *See id.* at 2.

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

115. *Model Multiplicity*, *supra* note 98, at 854.

116. *See* Benjamin Laufer, Manish Raghavan & Solon Barocas, *What Constitutes a Less Discriminatory Algorithm?*, 2025 ACM SYMP. ON COMPUT. SCI. & L., Mar. 2025, at 144 (using search strategies such as "sampling and re-training" to find LDAs).

117. *Less Discriminatory Algorithms*, *supra* note 114, at 70.

118. *See* Bruce E. Rosenblum, *Discriminatory Purpose and Disproportionate Impact: An Assessment After* Feeney, 79 COLUM. L. REV. 1376, 1392–93 (1979) (distinguishing the reasons for the use of foreseeability in the tort context from the goals of the Equal Protection Clause); Richard A. Warren, Personnel Administrator v. Feeney: *A Policy Decision*, 34 U. MIAMI L. REV. 343, 356 (1980) ("Left largely to its own discretion, the treatment of the evidentiary burden by the Court reflects more a policy decision than an interpretation of the law.").

chosen "in spite of" the predicted disparate results, the fact that a developer or policymaker could reliably predict such disparate impact (and decide whether to take steps to avoid it) or select a discriminatory model where an LDA exists implies a more active choice to discriminate.

2. AI Systems Generate a Limited Record for Reviewing Courts

There is also a practical argument for creating a carveout in the discriminatory intent analysis for AI systems: deploying AI models in decision-making systems might reduce the record available to reviewing courts. To identify discriminatory intent, courts often rely on the record produced by policy writers and implementers.[119] The need for a sufficient record is so great that the Court has expressly mandated that administrative agencies (those tasked with implementing a policy proposal) produce enough of a record for reviewing courts to evaluate.[120] With human decision-making systems, both policy writers and implementers create a record. Policy writers, when deciding how to frame a specific policy, generate a record of legislative history for how and why they settled on a particular solution. Policy implementers write explanations for how and why they reached a decision about a particular candidate. For example, judicial review of agency decision-making looks to whether an agency "examined 'the relevant data' and articulated 'a satisfactory explanation' for [its] decision, 'including a rational connection between the facts found and the choice made.'"[121] In contrast, AI models collapse the writing and implementation stages of policy development.[122] The policymakers and developers produce a record while creating the initial model, but once deployed, the AI model itself might not generate decision-specific explanations that would allow reviewing courts to identify impermissible discrimination.[123] In these situations there is less of a record for courts to evaluate during a discriminatory intent analysis. In other areas of law, to compensate for limited

119. McCreary Cnty. v. Am. C.L. Union of Ky., 545 U.S. 844, 862 (2005) (stating that courts consider "the traditional external signs that show up in the 'text, legislative history, and implementation of the statute'") (quoting Santa Fe Indep. Sch. Dist. v. Doe, 530 U.S. 290, 308 (2000)).

120. *See* Patrick M. Garry, *Judicial Review and the "Hard Look" Doctrine*, 7 NEV. L.J. 151, 166 (2006) (quoting Leslie Book, *The Collection Due Process Rights: A Misstep or a Step in the Right Direction?*, 41 HOUS. L. REV. 1145, 1198 (2004)); *see also* Citizens to Pres. Overton Park, Inc. v. Volpe, 401 U.S. 402, 419 (1971); Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43–44 (1983).

121. Dep't of Com. v. New York, 588 U.S. 752, 773 (2019).

122. *Cf.* Rashida Richardson, *Defining and Demystifying Automated Decision Systems*, 81 MD. L. REV. 785, 794, 809 (2022) (contrasting "[m]odern government decision-making," in which "decisions are delegated and distributed across multiple actors," with an algorithmic teacher evaluation system called a "Value-added Model" which is "criticized for being . . . opaque").

123. *See id.* at 809–10.

availability of evidence, courts reduce the evidentiary burden for showing a particular element.[124] Here, courts should do the same by compensating for the increased evidentiary difficulty of showing discriminatory intent by making it easier to find discriminatory intent. Yet even if an AI model produced decision-specific explanations, the Court would be ill-positioned to assess whether that machine-generated evidence is reliable.[125]

C. How to Incorporate Foreseeability into the Existing Doctrine

Courts can incorporate the foreseeability of disparate impact into their discriminatory intent analyses for AI systems in at least two ways. The first, and preferred, way is to infer discriminatory intent from a finding that disparate impact was foreseeable. The second, which would require minimal doctrinal change and is consistent with current Eleventh Circuit doctrine, is to include it as another factor in an *Arlington Heights* analysis.

The first approach treats foreseeable disparate impact for AI systems as sufficient grounds for inferring discriminatory intent. This proposed doctrinal change would create an explicit carveout to *Feeney* for AI systems. Under this approach, whenever developers' evaluations predict that a model will produce disparate impact,[126] developers have an obligation to search for LDAs — models with smaller between-group disparities but similar overall performance. If no search is conducted or an LDA is found and ignored, then courts would infer discriminatory intent because the post-deployment disparate impact reflects a choice by the policymakers to deploy a model that produces

---

124. For example, in tort cases where it is impossible to determine which tortfeasor caused the harm, courts effectively lower the burden of proof for the plaintiff by requiring the defendants to show which defendant caused the harm. *See* Summers v. Tice, 33 Cal.2d 80 (Cal. 1948) (finding joint liability for two defendants who both fired guns at the same time but where only one shot hit the plaintiff). Additionally, in negligence cases where access to evidence may be difficult, courts have implemented the principle of *res ipsa loquitor* to allow circumstantial evidence to infer negligence. *See* Judson v. Giant Powder Co., 107 Cal. 549 (Cal. 1895) (allowing a presumption of negligence following the explosion of a dynamite factory).

125. *See* Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 1983–89 (2017) (cataloging the "testimonial infirmities of machines," including "falsehood by design, inarticulateness, and analytical error," that collectively undermine the reliability of machine evidence).

126. This question is outside the scope of the paper, but other doctrinal areas might provide guidance about what counts as disparate impact. In the political apportionment context, the Court has generally used the baseline of at least ten percent deviation for discrimination cases. *See* Brown v. Thomson, 462 U.S. 835, 842 (1983). In the employment opportunity statutory discrimination context, courts have looked at different metrics such as "the impact ratio, the odds ratio, the flip-flop rule, the *Apsley v. Boeing* ratio, the standardized difference between disfavored and favored outcomes, and Cohen's h" to measure alleged discrimination. *See Practical Significance in EEO Analysis Frequently Asked Questions*, U.S. DEP'T OF LABOR, OFF. OF FED. CONT. COMPLIANCE PROGRAMS (Jan. 15, 2021), https://www.dol.gov/agencies/ofccp/faqs/practical-significance [https://perma.cc/2VAY-NLGB].

disparate impact despite the actual or likely availability of an LDA. But if no LDAs are found, then courts should not infer discriminatory intent because the developers looked for and failed to find less discriminatory policies, which suggests that the model creating the disparate impact was chosen "in spite of" its discriminatory consequences because there were no similarly performant options available.

The second approach differs from the first in that it would treat the failure to search for or adopt an LDA when there was predictable disparate impact as probative but not dispositive evidence of discriminatory intent. More concretely, for AI systems, courts would treat the foreseeability of disparate impact as a factor in an *Arlington Heights* analysis. And for more conservative members of the Court, this version of the test is consistent with existing doctrine. As the Court summarized in *Penick*, "[A]ctions having foreseeable and anticipated disparate impact are relevant evidence to prove the ultimate fact[:] forbidden purpose," as "the foreseeable effects standard" can "be[] utilized as one of the several kinds of proofs from which an inference of segregative intent may be properly drawn."[127] Consistent with *Penick*, some courts have already incorporated foreseeability into their *Arlington Heights* analysis. For example, in addition to the factors discussed in *Arlington Heights*, the Eleventh Circuit has also enumerated the following factors: "the foreseeability of the disparate impact; . . . knowledge of that impact, and . . . the availability of less discriminatory alternatives."[128]

Critics might respond that these doctrinal solutions focused on foreseeability would disincentivize developers from evaluating their models to avoid uncovering predicted disparate impact. But there are two constraints that mitigate this risk. First, because evaluating model performance is essential to the development process,[129] developers might prefer ensuring their models do not create disparate impact to removing such a commonplace part of the process. And second, conducting model evaluations and disclosing their results is either already mandated or could be mandated by state and federal legislation.[130]

---

127. Columbus Bd. of Educ. v. Penick, 443 U.S. 449, 464–65 (1979).

128. Greater Birmingham Ministries v. Sec'y of State for Ala., 992 F.3d 1299, 1321–22 (11th Cir. 2021) (citing Jean v. Nelson, 711 F.2d 1455, 1486 (11th Cir. 1983)).

129. *See* Jiang, Gradus & Rosellini, *supra* note 96, at 682 (explaining the role of test sets in validating model performance).

130. *See, e.g.*, S.B. 24-205, 2024 Gen. Assemb. (Colo. 2024) (mandating disclosure of evaluation methods and metrics).

IV. ALLEGHENY COUNTY FAMILY SCREENING TOOL CASE STUDY

A. Background on Allegheny Family Screening Tool

We use AFST as a representative case study for AI decision-making systems that rely on machine learning for several reasons. AFST's development process is well-documented, with explanations of the design, implementation, testing, and model selection process.[131] Additionally, the County documented its concerns about racial disparities and efficiency, allowing a transparent look into the motivations underlying AFST.[132] Finally, while prior academic work has discussed the Equal Protection implications of other algorithmic decision systems, such as those used in sentencing,[133] bail decisions,[134] or predictive policing,[135] there has been scant discussion of the Equal Protection implications of using algorithmic decision systems for child protective services.[136]

The County built AFST to help the Department of Human Services ("DHS") leverage the County's integrated data system to more efficiently protect children and investigate families for potential abuse.[137] Before AFST, when DHS received a child maltreatment referral implicating the General Protective Services statute,[138] the County hotline staff would have "discretion to respond by: (1) screening-out the allegation without further evaluation or assessment . . . (2) conducting a field screen of the maltreatment allegation in order to evaluate" whether

131. *See, e.g.*, *AFST Version 1*, *supra* note 64, at 5–33; *AFST Version 2*, *supra* note 4, at 2–9.

132. *Id.* at 6.

133. *See generally* Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 804 (2014) (arguing that including recidivism risk predictors that use demographic and socioeconomic variables to make bail decisions violates the Equal Protection Clause).

134. *See* Christopher Thomas & Antonio Pontón-Núñez, *Automating Judicial Discretion: How Algorithmic Risk Assessments in Pretrial Adjudications Violate Equal Protection Rights on the Basis of Race*, 40 MINN. J.L. & INEQ. 372, 387–88 (2022) (discussing how pretrial adjudications, such as bail determinations, may violate the Equal Protection Clause).

135. *See generally* Renata M. O'Donnell, Note, *Challenging Racist Predictive Policing Algorithms Under the Equal Protection Clause*, 94 N.Y.U. L. REV. 545 (2019) (arguing that machine learning-based predictive policing algorithms violate the Equal Protection Clause).

136. Prior literature focuses on the technical details or potential implications for racial equity, without extensive discussion of the Equal Protection implications of AFST or predictive modeling for child protective services. *See, e.g.*, Katherine Rittenhouse, Emily Putnam-Hornstein & Rhema Vaithianathan, *Algorithms, Humans, and Racial Disparities in Child Protective Services: Evidence from the Allegheny Family Screening Tool*, KATHERINE RITTENHOUSE 1 (Apr. 1, 2024), https://krittenh.github.io/katherine-rittenhouse.com/AFST_Disparities.pdf [https://perma.cc/HM9B-LC9K]; Gerchick, Jegede, Shah, Gutierrez, Beiers, Shemtov et al., *supra* note 12, at 1298–99.

137. *AFST Version 1*, *supra* note 64, at 4.

138. 23 PA. CONS. STAT. § 6334 (2014).

a full investigation was warranted, or "(3) conducting a formal investigation."[139] Though "historical and cross-sector administrative data" was already available to child welfare employees reviewing referral cases, the County feared that the data took too long to understand or was used inconsistently,[140] so it created an AI system that could aid "screeners in determining whether a [child] maltreatment referral [wa]s of sufficient concern to warrant an in-person investigation" into the home.[141] The County's stated goals for AFST were to "(1) improve the ability to make efficient and consistent data-driven service decisions based on County records, (2) ensure public sector resources were being equitably directed to the County's most vulnerable clients, and (3) promote improvements in the overall health, safety and well-being of County residents."[142] The County has launched multiple versions of AFST: "V1" in 2016, "V2" in 2018, and "V3" in 2022.[143] However, because the County has not yet published V3's implementation details, this Note focuses on V1 and V2.[144]

AFST's development process closely mirrored the process for other supervised machine learning systems. First, developers sifted through the available data sources to identify candidate variables that could serve as inputs into the model.[145] For both V1 and V2, the input variables were drawn from child welfare, jail, juvenile probation, and behavioral health records.[146] V1 also drew from public benefit records, while V2 instead used birth records and the allegations included in the child maltreatment referral itself.[147] Then, from this set of possible variables, developers used a feature selection algorithm to identify the subset of variables whose inclusion most contributed to the model's accuracy in predicting a child's "out-of-home placement within two

---

139. *AFST Version 1*, *supra* note 64, at 6 (emphasis omitted).

140. *Id.* at 5. *See also* Chouldecheva Benavides-Prado, Fialko & Vaithianathan , *supra* note 11, at 3 ("[T]he County has no means of ensuring that available information is consistently used or weighted by staff when making hot-line screening decisions . . . . [F]or example, recent paternal criminal justice involvement . . . may be a deciding factor in one case, while for another child with a similar referral that same information may be completely ignored.").

141. *AFST Version 1*, *supra* note 64 at 4.

142. *Id.*

143. Gerchick, Jegede, Shah, Gutierrez, Beiers, Shemtov et al., *supra* note 12, at 1310. *See also* Anjani Samant, Noam Shemtov, Kath Xu, Sophie Beiers, Marissa Gerchick, Ana Gutierrez et al., *The Allegheny Family Screening Tool's Estimation of Utility and Risk*, LOGIC(S) (Dec. 13, 2023), https://logicmag.io/policy/allegheny-family-screening-tools-overestimation-of-utility-and-risk/ [https://perma.cc/E2V7-VFM4].

144. As far as the authors are aware, there is no currently public documentation on V3, unlike the numerous publications from the developers of AFST for V1 and V2. *See Allegheny Family Screening Tool*, ALLEGHENY CNTY., https://www.alleghenycounty.us/Services/Human-Services-DHS/DHS-News-and-Events/Accomplishments-and-Innovations/Allegheny-Family-Screening-Tool [https://perma.cc/6E5P-QWVM].

145. *AFST Version 2*, *supra* note 4, at 3.

146. *Id.*

147. *Id.* at 3–4.

years of [a] screened-in referral."[148] With V1, the developers narrowed 800 potential variables to 71 highly predictive ones and used those variables to train a logistic regression model.[149] With V2, the developers combined the feature selection and model training steps.[150] Using the full set of 451 input variables, developers trained four supervised machine learning model types: LASSO, XG-BOOST, Random Forest, and SVM.[151] Only 126 of the 451 possible input variables actually contributed to V2's predictions.[152] Finally, V1 and V2 were validated using test set data pulled from actual records of removal.[153]

Concerns that AFST might "maintain[] or exacerbat[e] racial or socioeconomic disparities"[154] overshadowed the development process. Since AI decision-making systems have exacerbated existing societal inequities in other contexts,[155] researchers feared the same thing would happen with AFST, especially considering that "Allegheny County's historic data suggest[ed] that racial disparities already exist[ed] at many outcome and decision points throughout the child welfare system."[156] These concerns influenced the variables developers included as inputs and how they decided between models. When selecting input variables, the developers stated they would only include race as an input variable "if it substantively improved the predictive accuracy of the model."[157] Because a reported child's race did not improve model accuracy, it was not included as an input into AFST, which means the model is facially-neutral.

When deciding between models, developers prioritized (1) the "overall performance and accuracy for the specific high-risk group that serves as the focus of the County's policy" and (2) "equivalent levels

148. *Id.* at 10.

149. *AFST Version 1*, *supra* note 64, at 14, 33.

150. *See AFST Version 2*, *supra* note 4, at 10–11.

151. *Id.* at 6.

152. *Id.* at 10.

153. *Id.* at 7.

154. *Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation*, ALLEGHENY COUNTY (Jan. 2018), https://www.alleghenycounty.us/files/assets/county/v/1/services/dhs/documents/allegheny-family-screening-tool/allegheny-county-predictive-risk-modeling-tool-implementation-process-evaluation.pdf [https://perma.cc/3KTJ-A5ND].

155. Tim Dare & Eileen Gambrill, *Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County*, ALLEGHENY COUNTY ANALYTICS 6 (2017), https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf [https://perma.cc/X5YY-UUXL] (discussing the potential reinforcing considerations and comparing to the criminal justice context); *see generally* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016) (discussing ways in which data mining can exacerbate existing inequalities).

156. ALLEGHENY COUNTY, *supra* note 154 at 9 (citing MARY E. RAUKTIS & JULIE MCCRAE, THE ROLE OF RACE IN CHILD WELFARE SYSTEM INVOLVEMENT IN ALLEGHENY COUNTY, ALLEGHENY COUNTY DEPARTMENT OF HUMAN SERVICES (2010)).

157. *AFST Version 1*, *supra* note 64, at 15.

of accuracy for black children vs. non-black children."[158] AFST's developers measured performance using Area Under the Curve ("AUC"), which is a metric similar to accuracy that is calculated by taking the area under the curve created by plotting the true positive rate ("TPR") against the false positive rate ("FPR").[159] They narrowed in on high-risk groups by evaluating the positive predictive value ("PPV")[160] and TPR for high-risk referrals (those with referral scores greater than seventeen) and segmented across race to ensure that "the relative risk of placement for black children [wa]s similar to that for non-black children."[161] Ultimately, the developers selected the LASSO model because it had the smallest between-group performance disparities and was easier to implement.[162]

AFST is currently being used to generate risk-level predictions that assist child welfare employees in deciding whether to recommend a reported case of child maltreatment for further in-person investigation.[163] V2 added a new protocol where "the highest-risk cases [are] to be screened in" for further investigation by default, which can only be overridden by supervisors "with written justification."[164]

B. Applying the Foreseeability Analysis to the Allegheny Family Screening Tool

The foreseeability inquiry begins by determining whether an AI model's adoption produced disparate real-world effects. Precisely defining disparate impact for algorithmic systems is contested and outside the scope of this paper. But for clarity, we will provide a brief overview of some different approaches.

The first approach considers performance gaps between different demographic groups in the system deployed. These performance gaps can be measured either in accuracy or in the distribution of outcomes. Different evaluation standards correspond to different notions of

158. *AFST Version 2*, *supra* note 4, at 6.

159. *Id.* at 11. The TPR is "the proportion of all actual positives that were classified correctly as positives" and the FPR is "the proportion of all actual negatives that were classified incorrectly as positives, also known as the probability of false alarm." *Metrics Definitions*, *supra* note 102 (emphasis omitted).

160. The PPV is the likelihood that someone who has a positive result actually is positive. *See Positive Predictive Value*, NAT'L CANCER INST., https://www.cancer.gov/publications/dictionaries/cancer-terms/def/positive-predictive-value [https://perma.cc/KNN3-DCUE]. In the case of AFST, developers defined PPV as the "proportion of referrals flagged . . . where child ends up placed within 2 years." *AFST Version 2*, *supra* note 4, at 14.

161. *Id.* at 14.

162. *Id.* at 14–15.

163. *Id.* at 5.

164. *Id.* There is a similar override requirement for call workers requesting further screening in lowest-risk cases. *Id.*

fairness.[165] For example, a FPR is defined by how often a system incorrectly predicts someone as deserving the outcome.[166] Minimizing the FPR might be more "fair" in punitive contexts where we want to ensure no innocent person is incorrectly punished. This would accord with the Western maxim "that it is better that ten guilty persons escape than that one innocent suffer."[167] In the child-welfare context, optimizing for false negative rate (FNR) and accuracy might be more "fair." Similar to the false positive rate, the FNR describes how often a system incorrectly classifies someone as undeserving of a benefit.[168] Minimizing the FNR would reduce instances where a child's case is incorrectly screened-out (the case goes uninvestigated and the child is later hospitalized or injured). When allocating scarce resources, such as the time of child protection workers, higher system-level accuracy means resources are allocated more efficiently — more often allocated to (or withheld from) the people who need it (or do not). For AFST, an impact assessment compared "(1) the differential probability of screening in referrals involving Black vs. white children, and (2) the differential probability of home removal for referrals involving Black vs. white children."[169] The assessment found that "referrals involving Black children are more likely to be screened in than referrals involving white children,"[170] with "a 10.9 percentage point gap in screen-in rates before controlling for referral-level characteristics, which is reduced to 6.4 percentage points after controlling for referral-level characteristics and underlying risk scores."[171] This inquiry evaluates whether the impact of the policy alone could be disparate.

The second approach compares what existed before and after a policy was adopted.[172] Several impact evaluations of AFST assessed whether there were performance disparities between Black and non-Black or Black and White children by comparing the disparities in performance metrics, such as accuracy or false positive rates between groups, and also before and after AFST's deployment.[173] For example,

---

165. ACM FAccT Conference, *FAT* 2018 Translation Tutorial: 21 Definitions of Fairness and Their Politics*, at 16:57–18:02 (YouTube, Apr. 18, 2018), https://youtu.be/wqamrPkF5kk?t=1017 [https://perma.cc/4HMV-UY2R].

166. *Metrics Definitions*, *supra* note 102.

167. 4 William Blackstone, Commentaries on the Laws of England 358 (1791).

168. *Metrics Definitions*, *supra* note 102.

169. Rittenhouse, Putnam-Hornstein & Vaithianathan , *supra* note 136, at 4.

170. *Id.* at 17.

171. *Id.*

172. However, this approach has had mixed support from courts. *See supra* notes 34–35 and accompanying text.

173. *See, e.g.*, Jeremy D. Goldhaber-Fiebert & Lea Prince, Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County (Phase 2) 11–12 (2023); Rittenhouse, Putnam-Hornstein & Vaithianathan, *supra* note 136 at 3.

one assessment compared screen-in rates by risk score.[174] While there was no significant overall disparity reduction, for referrals with the highest risk scores of 19 or 20, AFST "significantly reduced disparities in screening decisions by 9.6 percentage points, or 98% of the pre-existing gap."[175] This upstream disparity reduction for screening decisions also mitigated performance disparities downstream, where, "among screened-in referrals, AFST reduced the disparity in case opening rates by 5.5 percentage points, (87% of the pre-existing gap) and reduced the disparity in 3-month removal rates by 2.9 percentage points (76% of the pre-existing gap)."[176] Another assessment estimated the accuracy of screening decisions by comparing short-term and long-term outcomes for screened-in and screened-out children.[177] By some metrics, V2 eliminated statistically significant disparities that existed before AFST's adoption.[178] But by others, such as the probability of another reported incident with a child who had been screened-out, the disparity widened.[179] That AFST increased performance disparities according to some metrics and decreased it according to others might lead a court to conclude there was no disparate impact. Ultimately, whether a particular policy change results in a disparate impact is still an open doctrinal question[180] that is especially complicated for AI decision systems[181] and outside the scope of this paper. For the rest of the inquiry, we assume that a court determined there was some impermissible disparate impact.

Assuming there was real-world disparate impact, the second part of the inquiry is whether that disparate impact was foreseeable by developers. This stage of the analysis looks at the similarity between the actual disparate impact and the predicted outcomes based on the test sets.

As part of AFST's pre-deployment testing, developers calculated the PPV and TPR for high-risk referrals, as well as the model's overall

---

174. *See generally Id.* at 14–15, 30 fig. 3 ("We plot screening decisions for low, middle and high risk scores, before and after the AFST deployment and separately for referrals involving Black vs. white children.").

175. *Id.* at 5.

176. *Id.*

177. *See* GOLDHABER-FIEBERT & PRINCE, *supra* note 173, at 1.

178. *Id.* at 20–21, 24 tbl. 2 (finding a reduction in disparity for GPS screen-ins that were Accepted for Service and GPS screen-ins that were accepted within 180 days).

179. *Id.* at 63–65 figs. D4–D6.

180. Coal. for TJ v. Fairfax Cnty. Sch. Bd., 218 L. Ed. 2d 71, 74 (2024) (Alito, J., dissenting from denial of certiorari) (discussing the Justices' disagreement about whether disparate impact should be measured by looking solely at the new distribution of benefits or by comparing the new distribution of benefits with the old one).

181. For example, disparate impact could be measured by considering disparities between the allocation of resources (e.g., how many people from each group are admitted to a school) or by considering disparities between model accuracy (e.g., is a model vastly less accurate for one group than another).

expected accuracy.[182] For PPV and TPR, the developers segmented the metrics by race to identify potential disparities and found that the "relative risk of placement for black children [was] similar to that of non-black children."[183] They ultimately selected the one with the smallest disparity, launching a model that had a relative risk of 3.10 of being placed for children flagged as high-risk and Black compared to those not flagged as high-risk and Black, compared to a relative risk of 3.20 for non-Black children.[184] This led developers to conclude that "conditional on race, the model[] [was] not miscalibrated."[185] For AUC, which is analogous to accuracy, the between-group disparity was larger. The model had an AUC of 74.42 percent for Black children and 77.35 percent for non-Black children, suggesting that AFST "was slightly better at predicting outcomes" for non-Black children.[186]

A court may not treat a 0.1 gap in relative risk or a three percentage point gap in AUC as disparate impact. The fact that the pre-deployment test sets did not find a disparate impact is unsurprising considering the post-deployment evaluations did not uncover real-world disparate impact. But one could imagine a counterfactual where the actual disparity was much larger than the predicted disparity — perhaps because the training data was old or unreliable. Whatever the reason, in such cases, the disparate impact would not be foreseeable because it was not known to developers in advance of deployment.

Although AFST's pre-deployment evaluations did not predict disparate impact, we again assume there was foreseeable disparate impact to illustrate the final step of our inquiry. This stage considers whether developers searched for and ignored an LDA. With AFST, the developers had at least three ways to generate alternative models that might be less discriminatory: (1) by varying the model algorithm types, (2) by changing the model parameters, and (3) by modifying the input variables.[187] The AFST developers tested four different algorithms for their model: Logistic Regression, LASSO, Random Forest, and XG Boost.[188] Within each algorithm type, they could also vary the model parameters.[189] Finally, they could change which input variables informed the model's predictions manually or algorithmically by

182. *AFST Version 2*, *supra* note 4, at 13–14, tbl.4 (describing the calculation and comparison of TPR, PPV, and AUC).

183. *Id.* at 14.

184. *Id.* at 14 tbl.5.

185. *Id.*

186. *Id.* at 7.

187. *See id.* at 10–11 (detailing the exploration of modeling techniques, such as the number of predictive variables used, the different models used, and their parameters).

188. *Id.*

189. *Id.* (discussing the potential variation of the number of "trees," "learning rate," "maximum depth of individual regression estimators," "regularization lambda," "regularization alpha," etc. for XG Boost).

adjusting the threshold for how much an input variable must contribute to a model's predictive power to be included in the final model.[190]

Although LASSO, Random Forest, and XG Boost were all similarly accurate,[191] LASSO had the smallest gap in accuracy between Black and non-Black children.[192] It also had the lowest relative risk disparity of 0.1, while Logistic Regression, Random Forest, and XG Boost had higher disparities, of 0.25, 0.58, and 0.79 respectively.[193] Partially because it had the smallest disparities along these two metrics, the developers ultimately chose the LASSO model.[194]

But consider a hypothetical in which the AFST developers chose the XG Boost model instead. Based on the pre-deployment test set results, they would know that it would likely result in greater disparities than the other models. In such a case, the decision to forego an LDA in favor of a more discriminatory model would be sufficient to infer discriminatory intent. One could also imagine a situation in which the developers had only trained the XG Boost model. If the pre-deployment evaluations suggested the model would produce a disparate impact, the developers would have an obligation to search for LDAs (e.g., the LASSO model). If they did not and the deployment of XG Boost resulted in disparate impact, then this would also support an inference of discriminatory intent.

V. CONCLUSION

Algorithmic decision-making systems like AFST enable new forms of discrimination, requiring new solutions from courts. Existing discriminatory intent jurisprudence is ill-suited to algorithmic contexts where intent can easily be hidden by proxies. In response to these algorithmic complications, we propose updating the discriminatory intent analysis to consider evidence from the results of test sets and the existence of LDAs that disparate impact was foreseeable and thus intentional. As state actors increasingly use algorithms to make some of the most consequential decisions affecting our lives, we hope this Note will help contribute to a deeper understanding of these systems' unique

190. *AFST Version 1*, *supra* note 64, at 13 n.8 ("Admittedly a t-ratio of 1.6 is rather arbitrary and based on judgment and experimentation with other cut-off levels.").

191. *AFST Version 2*, *supra* note 4, at 11 (explaining that LASSO achieved an AUC of 75.97, Random Forest achieved an AUC of 76.34, XG Boost achieved an AUC of 75.83, while Logistic Regression achieved an accuracy of 64.04).

192. *Id.* at 7.

193. These statistics were derived by subtracting the "Relative Risk of being placed if flagged as High-Risk and black vs. not flagged as High-Risk and black" from the "Relative Risk of being placed if flagged as High-Risk and non-black vs. not flagged as High-Risk and non-black," for each model. *See id.* at 14 tbl. 5.

194. *See id.* at 15 (describing how one important consideration for model choice was "equivalent levels of accuracy for black children vs. non-black children").

characteristics and inform a new line of Fourteenth Amendment jurisprudence that promotes more equitable and just algorithmic use.