

The Fair Lending Model: How the Longest-Running Algorithmic Fairness Programs Work in Practice

EMILY BLACK*, New York University, USA

MIRANDA BOGEN*, Center for Democracy & Technology, USA

LOGAN KOEPKE*, Upturn, USA

OLON BAROCAS, Microsoft Research, USA

WESLEY DENG, Carnegie Mellon University, USA

MINGWEI HSU, Upturn, USA

U.S. financial institutions subject to fair lending laws have been running algorithmic fairness programs for decades. Despite this long history, remarkably little is known about how these requirements operate in practice. In this paper, we offer the first empirical account of how financial institutions test for and mitigate algorithmic discrimination on the ground. In doing so, we shed light on how the regulatory design of fair lending law and regulation have shaped the policies, processes, and practices of fair lending programs. Drawing on 35 semi-structured interviews with participants across the fair lending ecosystem, we find that while financial institutions have a floor of fairness practices aimed at preventing discrimination in lending largely absent in other domains, the specifics of how firms test for discrimination and search for less discriminatory algorithms varies widely. We also find that regulatory supervision via fair lending examinations has been the key driver of compliance work, but that the practical impact of fair lending programs often depends on how well they can navigate competing business incentives, perceived legal tensions, and regulatory uncertainty. Ultimately, our findings highlight the unique role that supervisory authority has played in successfully fostering fair lending practices—a regulatory design feature that is distinct from other areas of civil rights law and almost completely absent from recent policy proposals for dealing with algorithmic discrimination.

CCS Concepts: • **Social and professional topics** → **Computing / technology policy**.

ACM Reference Format:

Emily Black, Miranda Bogen, Logan Koepke, Solon Barocas, Wesley Deng, and Mingwei Hsu. 2026. The Fair Lending Model: How the Longest-Running Algorithmic Fairness Programs Work in Practice. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT'26)*, June 27–30, 2026, Montréal, Canada. ACM, New York, NY, USA, 28 pages. <https://doi.org/10.1145/3805689.3812307>

1 Introduction

For over a decade, researchers have worked to draw attention to the harms resulting from algorithmic systems and have developed mechanisms to measure and mitigate a variety of these harms, including discrimination [2, 20, 21, 34, 46]. More recently, a line of empirical scholarship has begun to examine the challenges faced by industry practitioners working on responsible AI in practice to do so, particularly at large technology companies [31–33, 56].

*Equal contribution

Corresponding author: Emily Black, eblack@nyu.edu, NYU, New York, USA.



This work is licensed under a Creative Commons Attribution 4.0 International License.

FAccT '26, Montreal, QC, Canada

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2596-8/2026/06

<https://doi.org/10.1145/3805689.3812307>

Meanwhile, for nearly 50 years, U.S. financial institutions have been under a legal obligation to not discriminate in credit transactions [1, 75]. For the vast majority of that time, those same institutions have relied upon algorithmic decision-making systems to determine who gets access to credit and on what terms [12, 19, 63]. As a result, fair lending programs—programs to ensure that financial institutions comply with fair lending laws—are likely the longest-running example of algorithmic fairness on the ground. This overlooked history offers crucial lessons for the FAccT community, from both a technical and policy standpoint. Our paper is the first to empirically document the work of fair lending programs, and does so at a time of unprecedented erosion of fundamental building blocks of fair lending protection.

A range of scholars have offered a robust picture of potential methods to test for and mitigate discrimination and detailed the struggles practitioners face when attempting to do so [7, 33, 55, 66, 73, 87], but few of these studies have looked specifically at the role that formal regulations play in structuring those practices. The long history of fair lending programs offers a critical window into the way regulatory design choices—that is, how a regulatory system is structured [8]—can shape algorithmic fairness efforts in practice. Understanding how regulatory design can materially influence industry efforts to test for and mitigate discrimination is critically important, especially given the increasing focus on legal and regulatory mechanisms, tools, and instruments that may be necessary to advance algorithmic fairness.

Our study of fair lending programs at U.S. financial institutions focuses on three main questions:

RQ 1: How do fair lending teams at U.S. financial institutions test for and mitigate discrimination in algorithmic systems?

RQ 2: How does the regulatory design of fair lending law influence the policies, processes, and practices we observe in fair lending teams in U.S. financial institutions?

RQ 3: What challenges do fair lending practitioners face in ensuring that discrimination is effectively reduced?

To answer our research questions, we conducted 35 semi-structured interviews with participants across the U.S. financial industry, including engineers, lawyers, regulators, and third-party vendors. **Ultimately, this paper offers the first empirical account of how financial institutions test for and mitigate algorithmic discrimination on the ground. In doing so, we also shed light on how the regulatory design of fair lending law and regulation has shaped the policies, processes, and practices of fair lending teams.**

For efforts to advance algorithmic fairness to be successful at both a methodological and structural level, we argue they must be rooted in evidence on how regulatory design impacts algorithmic fairness work on-the-ground. Such evidence is of particular importance now, as policymakers across the country are introducing both legislative and regulatory proposals with the aim of better combating algorithmic discrimination [10, 18, 25, 40, 57, 58, 74]. Ultimately, we believe that by understanding what firms have done—and specifically by understanding how those efforts were informed and shaped by specific regulatory design choices—advocates, practitioners, policymakers, and scholars will be better positioned to understand what regulatory conditions are needed to reduce algorithmic discrimination and why. While we set out to elicit insights that could support technical and policy efforts to reduce discrimination in other consequential contexts, our findings ultimately also serve to document the financial industry’s practices immediately leading up to the current U.S. administration’s efforts to actively dismantle the very regulatory regime of interest.

2 Background

Four laws collectively serve as the foundation for fair lending regulation in the U.S.: The Equal Credit Opportunity Act (ECOA), the Fair Housing Act (FHA), the Home Mortgage Disclosure Act (HMDA) and the Community Reinvestment Act (CRA) [1, 75–77]. Passed in 1968, the FHA prohibits discrimination when renting or buying a home, getting a mortgage, or seeking housing assistance [1]. ECOA (1974) forbids discrimination in any aspect of a credit transaction on the basis of race, sex, age, and other protected characteristics [75]. Both the FHA and ECOA

prohibit disparate treatment, which emphasizes discriminatory intent, and practices that have a disparate impact.¹ HMDA (1975) requires financial institutions to maintain, report, and publicly disclose loan-level information about mortgages, in order to help identify potential discriminatory lending patterns [76]. The CRA (1977) is designed to ensure that mortgage lenders serve low- and moderate-income neighborhoods, thereby combating practices like redlining [77]. Notably, only ECOA and the FHA impose substantive legal obligations on financial institutions to not discriminate. HMDA, meanwhile, compels the production of data that plaintiffs and regulators can rely upon to identify discriminatory patterns, whereas the CRA imposes a general obligation for lenders to meet the needs of their community.

The regulatory design of fair lending law is strikingly different from other civil rights laws. By regulatory design, we mean the ways in which a regulatory system is structured—from the institutional form of the regulator to its enforcement tools to its compliance mechanisms—to achieve its stated goals [8, 23]. To understand why, consider the main ways in which Title VII, the foundational federal law prohibiting employment discrimination, is enforced. First, aggrieved individuals who believe they have been victims of discrimination must file a complaint with the Equal Employment Opportunity Commission (EEOC) [36]. The EEOC will investigate these complaints and either file their own lawsuit or issue a Notice of Right to Sue, allowing the aggrieved individual to vindicate their private right of action and file suit against an employer for discrimination [36]. The EEOC is almost completely reactive: to enforce Title VII, it almost exclusively relies upon aggrieved individuals filing a complaint of discrimination [79].

Fair lending laws like the FHA and ECOA also equip victims of discrimination with a private right of action to sue lenders for discriminatory actions and empower regulators to investigate complaints. However, financial regulators who enforce fair lending laws—like the Consumer Financial Protection Bureau (CFPB)—also possess robust statutory authorities that authorize them to proactively and routinely inspect business practices for compliance with fair lending laws, *whether or not the regulator believes discriminatory practices are occurring* [79]. Such supervisory authority and examination, where regulators actively oversee regulated entities' compliance by sending examiners to a financial institution's offices to probe for evidence of lending discrimination [24], are the key to the regulatory design of fair lending law. Through supervision and examination, fair lending regulators can direct financial institutions to address potential deficiencies in antidiscrimination efforts. If such deficiencies are not resolved, they can be elevated to formal enforcement proceedings. Ultimately, "supervision is where the practice of government actually occurs in finance" [28].

Finally, this regulatory regime only applies to legally defined "covered entities." These include institutions that directly furnish credit to consumers, but does not include, for instance, vendors of credit scores [71]. In the remainder of the paper, we refer to entities as either "covered entities" or "entities not covered" to indicate whether they are subject to these laws. We note that some elements of these laws, regulatory regimes, and supervisory priorities are in active flux in response to executive actions under the current U.S. administration.

3 Related Work

A growing body of research in fields such as Human-Computer Interaction (HCI) and Responsible AI (RAI) has investigated how industry practitioners engage in fairness work within real-world organizational contexts [5, 30, 31, 49, 55, 56, 61, 62, 64, 72, 81, 82]. Several works have revealed how practitioners evaluate model fairness in their organizations, identifying technical challenges in bias mitigation in practice, like selecting appropriate metrics and determining which stakeholder groups to prioritize [56], and studying use and misuse of popular open-source fairness tools [7, 32, 72]. More generally, many have emphasized the difficulties in concretizing

¹Since data collection, the CFPB finalized a rule removing disparate impact liability from Regulation B, the regulation implementing ECOA. This changes the agency's regulatory posture but does not resolve whether ECOA itself prohibits disparate impact, which courts have consistently found it does.

abstract responsible AI best practices or principles, suggesting these soft requirements may not often lead to more responsible systems in practice [3, 5, 17, 44, 54, 55, 67, 73, 83]. While a few works have touched on the impacts of regulatory requirements around protecting user privacy [53, 69], none have examined fairness practices in organizations explicitly covered by civil rights law, which has required the mitigation of unlawful disparities in credit, housing, and employment contexts for decades. Similarly, while others have provided a law and policy overview of fair lending considerations for machine learning models and algorithmic fairness [45, 51], these works do not detail what financial institutions do in practice. More recent work outside the academy has surveyed relevant fair lending techniques [37, 38], but has not considered how regulatory design has shaped compliance practices.

An emerging body of work within the FAccT community has sought to evaluate the efficacy of new regulations targeting algorithmic discrimination in employment, specifically New York City's Local Law 144 [39, 42, 52, 85]. These studies have largely focused on establishing the degree to which covered actors appear to be complying with the law, assessing compliance quality, and identifying weaknesses in the law that might account for its lack of efficacy. Unlike our study, these works focus on newly implemented regulations, rather than well-established civil rights laws.

Additionally, little prior work has directly engaged compliance practitioners as study participants, in part because many existing studies have not focused on highly regulated domains like the financial sector, where compliance requirements can strongly shape fairness testing and mitigation practices [4, 22, 59]. Many prior studies have acknowledged as a limitation the challenge of recruiting compliance professionals, who, given the sensitivity of their work, are often harder to access than technical practitioners [33, 55, 68, 80]. Our work extends the literature by being among the first to investigate how fairness testing and mitigation are carried out in practice in the highly regulated domain of lending, extending empirical insights beyond engineers or researchers to include lawyers, third-party vendors, and regulators.

Furthermore, while many empirical studies of RAI in practice highlight regulation as a key lever for motivating and scaffolding fairness work [31, 61, 64, 65, 84], the FAccT community still lacks an in-depth empirical understanding of how regulatory design can shape firms' efforts and practitioners' work. This gap is especially consequential for fairness testing, where existing policies and regulatory guidance are frequently described as incomplete, difficult to operationalize, or misaligned with practitioners' day-to-day workflows [5, 33, 54, 55]. Through an in-depth interview study with practitioners across roles, our work offers concrete insights about how details of regulatory design can significantly shape the antidiscrimination testing done—and not done—inside corporations. Ultimately, our work is perhaps most in line with prior sociolegal studies of firms seeking to comply with discrimination and privacy law [9, 35, 78], which find that firms are able to take advantage of significant latitude in how to interpret regulation—and often do so in ways that substitute performative compliance practices for meaningful changes in corporate behavior.

4 Methods

Study Design. We conducted semi-structured interviews with 35 participants related to financial services in the U.S. We developed different interview protocols for our four main populations: engineers, lawyers, third parties, and regulators. We first conducted a pilot study with six participants to help construct effective protocols for each role. Once protocols were in place (see Appendix B.1, B.2, and B.3), we conducted semi-structured interviews to understand participants' current practices and challenges around testing and mitigating discrimination in algorithmic systems. Semi-structured interviews each lasted up to an hour. The interview began by asking participants to describe their current role, their previous relevant experiences, and their organization's process and structure as it relates to algorithmic fairness. Then, we tailored questions depending on the population.

For example, we asked engineers about techniques and metrics, while we asked lawyers how disparate impact doctrine informs fairness testing in their organization.

Participants. We adopted a purposive sampling approach with the goal of recruiting practitioners across multiple roles who had experience testing and mitigating discrimination in algorithmic systems broadly deployed in the financial services sector. We reached out to more than 100 practitioners through a combination of direct outreach and snowball sampling, leveraging the research team’s professional networks and referrals. Ultimately, 35 participants, spanning dozens of organizations, took part in the study. Table 1 in the Appendix provides an overview of these participants. We note that we refer to participants as either E, L, T, or R depending on whether their role matched most closely with an engineer, lawyer, third-party actor, or regulator.

Given the sensitivity of fair lending and compliance-related work, we took several steps to protect participant confidentiality. Following prior work on responsible AI practices in industry, we omitted demographic details and abstracted certain information about participants’ employers and roles to avoid inadvertently identifying individuals working in this high-stakes domain. In addition, we assured participants that we would not ask them to reveal any confidential or personally identifying information about themselves or their colleagues and that we would de-identify all responses at the individual, team, and organization levels. All participants were offered a \$50 honorarium for their time, though many declined due to organizational policies. Our protocol was approved by our institutions’ IRB.

Data Analysis. Our study sessions yielded approximately 38 hours of audio that we transcribed. To analyze our interview transcripts, we adopted a reflexive thematic analysis approach [15]. Members of our research team met after each interview session to conduct an interpretation session and collaboratively develop a codebook for annotation, which we include in the Appendix. We then distributed the transcripts among the research team for open coding. Each transcript was analyzed by one team member, typically the person who led the interview. Another team member then reviewed and validated the interpretations made by the primary annotator. Throughout the coding process, the authors regularly discussed discrepancies in interpretation and iteratively refined the codes based on these discussions. Consistent with reflexive thematic analysis, we did not calculate intercoder reliability, as coding is understood to be interpretive and iterative rather than aimed at statistical agreement [15]. In total, we generated around 1,200 unique interpretation codes. Through an iterative, bottom-up affinity diagramming process, we grouped our interpretations into successively higher-level themes. We present our results in the following section, organized around our three research questions.

5 Findings

5.1 RQ1: The contours of fair lending practice

In this section, we describe the overall structure of discrimination testing and mitigation procedures by regulated financial institutions as explained by our participants. While participants broadly agreed that firms are expected to test for and prevent discrimination, and our data suggest that fair lending teams and formal antidiscrimination procedures are commonplace, the specific practices employed vary widely.

5.1.1 Existence of Fair Lending Teams and Practices. As a first order matter, our interviews showed that fair lending teams invariably *exist* within regulated financial institutions—which, as prior work has noted, is not always the case at firms that are not covered by fair lending law [56]. All of the individuals we interviewed who were at institutions covered by fair lending law described a designated fair lending unit that was responsible for monitoring compliance with relevant laws ($L_5, E_1, E_6, E_4, L_4, E_{11}, L_2$). As L_5 , who worked in such a department, described, “*it is my responsibility to make sure that our bank does not discriminate against any of our customers or potential customers.*”

Beyond the existence of teams dedicated to antidiscrimination work, our interviews revealed a ubiquity and formality of fair lending procedures. Participants at regulated financial institutions almost universally believed

they were expected to test models for potential discrimination, and as a result, had procedures to do so (e.g. L_2 , T_1 , L_3 , E_4 , T_5 , E_6 , R_1 , L_5 , E_8 , E_9). As L_6 stated, “*you need to test*” lending models for discrimination. Participants connected these procedures to regulatory requirements: L_2 said “*The [Consumer Financial Protection] Bureau... [has] been clear that it’s a responsibility of the lender to make sure that their models are not causing disparate impact.*”²

While specific testing procedures varied between firms, a required procedure always existed, with many common elements. Many participants at regulated financial entities stated that a model would first be triaged to determine what level of bias testing would be pursued. For example, L_3 shared that “*we have a policy that essentially allows for a model to be risk rated, depending on how it’s being used and what it does, and then a testing process that is basically aligned with that risk rating.*” The risk of a model is often related to the probability that its outputs would impact a consumer: T_9 shared, “*the closer the model is to the consumer, the higher the fair lending risk.*” E_9 noted that firms would conduct “*a much deeper dive*” on models with the highest fair lending risk versus “*a little bit less in-depth analysis*” on other models.

5.1.2 Procedures to Test for Disparate Treatment and Disparate Impact. In order to prevent disparate treatment, participants (L_2 , T_1 , L_3 , E_4 , T_5 , E_6 , R_1 , L_5 , E_8 , E_9) detailed multiple procedures to review individual model features—that is, tests to ensure that none of the input features to a model were themselves or served as an obvious proxy for a protected characteristic such as race. For example, T_1 described how most firms will “*start with a qualitative variable review*” involving a review of a model’s features not only for protected characteristics (meaning, prohibited bases protected under state and federal law), but also whether the model’s features were “*intuitively related to credit worthiness.*” Quantitative feature review would then involve assessing whether other features, even if not explicit prohibited bases themselves, were proxies that are “*highly correlated with a prohibited basis*” (L_3). As part of both quantitative and qualitative feature review processes, R_1 observed that institutions often develop “*green and yellow and red lists for variables*” where red features are “*not allowed in models,*” yellow warrant further analysis, and green have a clear and defensible relationship to credit.

Models that firms identify as posing the most fair lending risk—such as underwriting or pricing models—would then also be subject to a disparate impact review (L_2 , T_1 , T_5 , T_8 , L_7 , T_9 , T_2): as L_7 shared, “*our highest [model] risk category [...] gets the deepest level of testing: a full disparate impact analysis.*” As we discuss in more detail in the next section, we observed variation in how institutions defined disparate impact, as well as the thresholds at which a disparate impact triggered further analysis. Participants (T_1 , T_5 , T_8) explained that disparate impact testing relied on “*tests like adverse impact ratios [(AIR)], standardized mean differences [(SMD)], [and] marginal effects*” (T_8), with a focus on determining whether disparities were statistically and practically significant. Other participants examined disparities not at the model level in abstraction, but instead through the lens of historic outcomes. L_7 ’s institution would look at the default rates for various protected classes (e.g., the historic default rate for black and white borrowers) and then compare a model’s predictions against those historic outcomes to see if the model would worsen those disparities. If the model did not appear to worsen outcomes relative to historic rates, L_7 described that “*we consider the model to have low fair lending risk.*” In other words, if the predicted default rate for specific subpopulations was in line with their historical base rate of default, the institution believed that it had a satisfactory business justification for the disparate impact.

Participants believed this variance was in part due to a lack of clear regulatory expectations on how such tests should be performed. L_2 noted there are “*still no metrics or no concrete, crispy, guidance*” from regulators. E_9 concurred, stating “*there is no regulatory guidance that this is the exact way how you should be doing this disparity testing. So as long as it’s reasonable [...] there is no black and white, right or wrong, answer here.*” Conversely, clear

²In fair lending law, companies must prevent both disparate treatment (they may not make lending decisions on the basis of protected characteristics such as race) and disparate impact (they may not have underwriting procedures that result in disparities unless they can be justified by business necessity) [75].

regulatory expectations regarding methods to test for discrimination resulted in some standardization across industry, and a floor of practice. For example, the CFPB's whitepaper [26] indicating that Bayesian Improved Surname Geocoding (BISG) could be used to impute race information needed to perform discrimination testing offered many companies a tacit path to conducting discrimination testing despite institutions' trepidation around the collection of actual demographic data from customers.³ Many participants ($T_8, R_1, R_2, T_5, T_3, L_3, T_1, E_6$) described industry use of BISG as a standard. As T_1 summarized, "*I think that the CFPB issuing that BISG paper went a long way*" to making a free, easy, and public method supporting disparity measurement widely available, explaining that "*it's not perfect, but we can all coalesce around [it]*." This contrasts with many studies of RAI in industry, which observe how constrained demographic data access can create friction for discrimination testing [33, 49, 53, 69].

5.1.3 Searching for Less Discriminatory Algorithms. Following testing, many participants stated that if a model was found to exhibit disparate impact according to their firms' metrics, they would search for a less discriminatory algorithm (LDA)—that is, a model achieving a similar performance with less disparity ($L_2, L_3, E_1, E_4, E_6, E_{11}, E_9$). As we discuss in Section 5.3.2, LDA search methods, and thus their probability of success, varied significantly across companies. Most participants reported *drop-one* analysis [37]—removing one input feature at a time and retraining—as their primary bias mitigation method. However, a few participants (e.g., E_1) described more robust search processes that involved developing thousands of candidate models with different design choices and selecting the fairest among them within a 1-2% performance tolerance (T_5, E_{11}).

5.2 RQ2: Driving forces of discrimination work in regulated financial institutions

We also engaged participants on the motivations that drove their antidiscrimination work, in order to understand how the *regulatory design* of fair lending law structured their practice. In particular, we aimed to identify factors in the implementation and enforcement of fair lending law that seemed to specifically motivate institutions' policies, processes, and practices, as well as how they made compliance-related decisions. As a first order matter, our interviews revealed that a floor of practices exist in heavily regulated institutions that do not seem to reliably exist in entities not covered, suggesting that the mere existence of fair lending regulation is a driving force for antidiscrimination work. As L_7 summed up, "*if it wasn't for the fair lending legal risk, we wouldn't even do this review.*" More specifically, our participants shared that regulatory expectations and scrutiny were the main motivators behind the development and execution of fair lending work, as opposed to individual lawsuits brought through private right of action.

5.2.1 Participants cite regulatory expectations as a reason why their work was done at all. In our interviews, participants described how they would use regulatory expectations and fear of enforcement to internally motivate their work (L_2, T_5): as T_5 shared, for example, "*Unfortunately, no one does this stuff out of the good and kindness of their heart. They do it because they're afraid of regulation.*" L_2 expanded, noting how they would use enforcement actions to further motivate internal work, stating that "*All the time, the fair lending team at [company name] would say, hey, this settlement just came out; you know we're doing pretty well in this, but we need to raise the bar because the settlement came out.*" Conversely, participants shared that without regulatory expectations, antidiscrimination work would simply not happen (R_4, L_1). As T_{10} stated about the domain of insurance, where they did not observe regular bias testing or mitigation, "*They have to be required to do it in order for them to do it [...] they're not gonna do it otherwise, in my opinion.*" As a further disincentive, L_1 explained that testing without a legal mandate could actually introduce additional legal risk to a firm.

³Some of this trepidation may stem from the law itself, which explicitly *requires* discrimination testing when demographic data are collected [75].

5.2.2 Supervisory expectations and fair lending examinations as a primary motivator. While the general existence of fair lending law motivated firms' antidiscrimination efforts, participants were clear that supervisory expectations and fair lending examinations by regulators were the most consequential factors driving specific fair lending work at their firms, not private litigation. For example, L_7 directly stated that “we spend more of our time worrying about [the] regulators, because we go through fair lending exams all the time.”

Regulatory scrutiny not only drove whether work was done, but also what specific work was done: L_7 suggested that “the regulators expect us to be searching for less discriminatory alternatives and documenting our searches,” continuing to explain that their search process was “well documented so that we can sort of show we’ve done our homework.” L_7 ultimately concluded that, “if I was thinking about the motivations for why we take that action, it’s really probably more to avoid an unpleasant fair lending exam than it is to worry too much about the litigation risk.”

Regulatory expectations also shape what work was *not* done. For example, E_6 noted that fair lending testing of credit models mostly meant testing for disparate impact on a limited set of protected characteristics (race and sex), despite the law covering multiple other protected classes: firms were “not very interested” in figuring out how to do tests for other protected characteristics “because from their perspective [...] the regulators are fine with what we’re doing.”

5.3 RQ3: Challenges to fair lending work in practice

Despite well-established fairness testing and mitigation processes, participants still described many challenges that complicated their ability to reduce disparate impacts. Here, we highlight three main challenges participants surfaced: first, the reality that the decision to follow the advice of the fair lending team is a business decision; second, that perceived legal tensions led to technical difficulties in mitigating disparities; and third, that uncertainty about regulatory requirements could be leveraged to justify the bare minimum to prevent discrimination. At times, though, this uncertainty seemed to provide some fair lending teams the flexibility to adopt surprisingly progressive policies and practices.

5.3.1 Frustration with compliance being a business decision. Participants shared that the first challenge in avoiding the use of biased models in practice was convincing the business arm of the organization to do so. They explained that deciding whether to take on the risk of deploying a model resulting in a disparity is in the hands of the business arm of an organization, which weighs the cost of compliance against the cost that would be incurred by a potential violation and the upside of proceeding, and decides whether to reduce the disparity. As E_6 described: “fair lending is a legal advisor” but the business “can choose to take the advice and ignore it [...] it is definitely within policy for an executive [to] decide that [an LDA] is not worth the cost, and they’re gonna go ahead with [the original model], but then they bet the entire legal risk of what it was to go ahead with.” Several participants described this dynamic as well (L_2 , L_3 , T_8 , E_{11} , L_5 , E_9). As L_5 stated, there is little to no obligation to sacrifice some business objective in order to reduce disparity: “I don’t think there’s an expectation that we have to make [...] financial decisions that are money losers. I don’t think we have to make unprofitable decisions.”

Beyond deciding whether or not to deploy an LDA, participants shared that the business arm of their firms were also in charge of ultimately deciding what counted as a viable LDA—in other words, how much leeway in model performance to grant in order to facilitate disparity reductions. The process of defining performance or cost thresholds for LDA viability was nearly always ad hoc, so decisions and practices varied widely: as E_4 observed, employees charged with model risk management (largely focused on “safety and soundness,” or model performance, who might prefer minimal model degradation) and employees charged with fair lending (focused on reducing discrimination, and open to some minor model degradation if the alternate model resulted in meaningful narrowing of outcome gaps) would have to “hash it out.” These conversations, and the need to constantly advocate for less discriminatory options, were a reported source of frustration for participants in fair

lending teams: for example, T_5 shared that this question was “*the source of half of my misery in life*.” T_8 shared that “*there’s a lot of conflict*” in these conversations.

Many participants reported that their firms had incredibly low tolerance for bearing performance costs to their models in order to reduce disparity: L_2 noted that “*modeling teams would be like no, the [LDA model] has to perform equally as well or we’re not going to make the change*,” where “equally as well” was interpreted very strictly. Similarly, T_9 noted that some modelers, when looking at a baseline model with a KS^4 of .5, and alternatives with a KS of .4, .48, .49, would suggest that “*unless I get .5, it’s not a viable alternative*.” T_5 shared that modelers they worked with would often make the argument that “*a .000001 change in the AUC is meaningfully different*.”⁵ We note that the way fair lending teams measured model performance, and what metrics of performance they traded off with disparity, also varied widely, with some using AUC, others using KS , others simply using accuracy, and still others forecasting directly to dollars lost.

Notably, participants (T_5 , T_9) observed a level of internal inconsistency between model degradation tolerated for other model risk management processes and for fair lending, which they thought was unjust. As T_9 said: “*If you allow your model, before you have to [retrain to deal with drift], to degrade by 2 and a half percent, 5 percent, why can’t you allow that [to combat disparate impact]? If it’s not meaningful for you from a safety and soundness standpoint, why is it meaningful for you from a [disparate impact] standpoint?*” Some participants suggested some of this conflict could be the result of “*cultural issues*” with data scientists, stemming from how they “*were trained*.” (T_8).

However, most participants believed that the lack of uniformity here was in part due to a lack of regulatory clarity around an acceptable threshold—and some specifically noted they were eager for one (E_{11} , L_6 , E_8 , L_7). As E_8 shared, “*My job will be easier if tomorrow the CFPB comes with prescriptive guidance [...] with a number*.” However, as we see in Section 5.3.3, finding the line between clarity and specificity can be a difficult task. And, as we discuss in the next section, clarity in regulatory expectations can only advance work so far when perceived legal tensions linger.

5.3.2 Perceived tensions: disparate treatment and impact. Fair lending programs must manage risks related to two doctrines in discrimination law, disparate treatment and disparate impact: while disparate treatment outlaws decision-making *on the basis of protected characteristics* such as race, disparate impact aims to reduce unjustified disparities in decision-making outcomes. Participants most often mentioned challenges navigating both theories simultaneously when discussing how their organizations think about mitigating gaps revealed by disparate impact analyses, particularly with regard to the extent to which demographic information may be considered when searching for LDAs.

A discussion of the legality of algorithmic bias mitigation techniques in light of disparate treatment constraints is out of scope of this paper (for a detailed treatment of this issue, see [50]). Nonetheless, our research underscored the practical impact of the *perception* that a direct tension between these two legal doctrines exists, especially in the case of AI systems, within the fair lending community. Participants’ reflections revealed how this perceived tension influenced companies’ bias mitigation procedures, which they observed regulators have generally found to be acceptable. However, even as many practitioners find these procedural design choices to be increasingly outdated in light of more complex AI systems, they observe that their firms have been reticent to update approaches without an explicit go-ahead from the regulator. At the same time, regulators were aware that while they could use their discretion to communicate expectations, ultimately they alone cannot resolve this sort of contentious topic.

⁴The Kolmogorov-Smirnov (KS) metric measures the maximum separation of the cumulative distribution functions of the positive and negative classes of a model’s predictions, i.e. defaulters and non-defaulters. This metric demonstrates the model’s ability to meaningfully distinguish between classes.

⁵The Area Under the Receiver Operating Characteristic (AUC) curve plots the true positive rate against the false positive rate for all possible thresholds.

Participant frustration with seemingly inefficient institutional design choices. One popular strategy financial institutions employ to navigate the perceived tension between disparate treatment and impact is a complete separation of individuals working on building financial models subject to fair lending law (the “first line of defense” in firms’ risk management structure) and those investigating them for, and potentially mitigating, bias (the “second line”). This separation typically dictated access to demographic data: many participants shared that only the second line would have access to demographic data necessary to conduct fairness tests. E_4 explained, “*you never gave the modelers access to the fair lending attributes [since you] ...didn’t want anybody putting race or ethnicity, or anything like that in a model.*”

Many participants confirmed this set-up (E_1, E_4, E_6, L_6, T_9): for example, E_4 noted that “*modelers did not gain access to protected class data.*” This firewall also prevented first line modelers from even seeing the results of the testing done by the second line (E_1, E_6, E_8): For example, E_6 observed that if the first line “*got a negative outcome from the disparate impact analysis,*” they weren’t “*allowed to see*” the disparate impact analysis “*so they would just be told to change the model, but then [they] wouldn’t know how to change it.*”

While this organizational practice may have been a plausible way to manage discrimination risks in a paradigm of simpler, smaller models for which new versions could be developed with reasonable ease, participants shared that such structures serve as a serious impediment to effective mitigation when dealing with today’s more complex, larger models (E_1, E_6, L_6, T_9). As a result of the separation and lack of communication between the two teams, E_6 described an inefficient “*back and forth, trying to figure out how to fix the model without having access to analysis.*” T_9 agreed, stating that the first line/second line separation meant that “*you need to have two complete new model teams. It just doesn’t really work well.*” L_6 expressed that in “*an ideal world,*” the first line would do the bias testing alongside the model development, since explaining how the model works well enough to another team for them to test it takes “*a lot of time.*” Several participants expressed deep frustration with this set-up, some even leaving their jobs over the practice. As E_1 explained, “*[D]on’t you want our data scientists to understand fairness so that we are engineering it in at the point of model development rather than doing it after the fact? And so, it took a very long time for [the second line] to even warm up to the idea that they should even talk to me as representing part of the business. This is something that honestly I was not able to successfully move the needle on very much at [company name], which is also part of the reason why I left.*”

Some participants observed their companies starting to deviate from this structure (E_{11}, L_6, T_9). As L_6 noted, some firms have developed a process where the first line develops the best performing model, “*draws a line,*” and then performs disparate impact analysis. That is, they separate the two processes within one team, to preserve independence of the processes while reducing information flow issues. T_9 agreed that firms are moving to establish a “*clear delineation as to who in the first line is going to do the [disparate impact] work [...] I think that’s where we’re all kind of landing right now.*”

However, progress towards this change appears to be slow, and marked by continued concerns over implicating disparate treatment. For example, L_7 described the process at their institution, where first line developers did have access to some basic disparate impact information through an “*automatic bias estimator [...] where the modelers can actually do a little bit of [disparate impact analysis] during the development process without necessarily giving the modelers access to race and ethnicity data.*” Critically, though, L_7 noted that their organization places limits on how many times modelers can use the automatic bias estimator: “*You could in theory [...] run iterations through 150 different times, and basically snuff out [...] it sort of becomes a backhanded way of getting to the demographic information,*” and that in response they impose a limit on “*how many runs you can put through the thing, to specifically avoid that problem.*” Some viewed such approaches as a practical path forward, allowing LDAs to be surfaced earlier in model development while mitigating disparate treatment concerns by not revealing raw demographic information to models or modelers.

Frustration with reliance on out-of-date mitigation methods. Participants described that, in practice, many methods that have been proposed by researchers to mitigate discrimination in AI systems are seen as inducing a risk of disparate treatment discrimination [48, 50].⁶ This seemed to be driven by firms' interpretations of disparate treatment—which prohibits decision-making directly on the basis of a protected attribute such as race—to mean that protected class information should not be used at all during the course of model development. As L_5 shared, “*Those of us in the industry don't think we can actually factor a person's demographics into [...] the building of a model in order to lessen discrimination, at least not factored in directly [...] Whereas there are [...] academics or others outside of the industry that feel like that is a way mathematically that can be helpful.*” However, much of the algorithmic fairness literature assumes at least some access to demographic data, and model debiasing methods informed by demographic information are understood to be more effective at reducing algorithmic discrimination than anticlassification (i.e. fairness through unawareness). Perceived legal risk has led many financial institutions to opt not to use these more modern debiasing methods. As T_9 explained, while regulators in recent years expressed an expectation that firms be proactive in mitigating bias, they “*haven't fully signed off on*” many algorithmic debiasing methods financial institutions may want to use to reduce disparate impact. E_6 also stated that their fair lending department was resistant to deviating from the status quo by amending legacy methods: when they tried to collaborate with their firm's fair lending team, E_6 found that fair lending insisted that their “*methods are good enough, and they [fair lending] get very defensive [...] the paradox is that you can't improve procedures if your position is that legally, we are doing everything we need to do.*” Indeed, for many participants, basic methods like drop-one analysis were the main or only debiasing method discussed during our interviews (T_2, R_1, T_9, E_9). R_1 , a former regulator, affirmed that in general, drop-one or variable changes “*tends to be as far as companies go.*”

Despite this inertia, participants shared an overall recognition that such approaches were no longer effective for more complicated AI systems. Whether or not their firms permitted use of newer methods from the algorithmic fairness literature, most of our participants believed that such methods would be necessary to actually tackle biases in AI systems (E_4, E_6, E_{11}, E_9). For example, E_{11} stated that “*drop one is an old school method that that doesn't work anymore.*” Thus, if a company is not willing to experiment with newer methods due to fears of implicating disparate treatment—which many were not—our data suggests that this perceived tension led to a much less effective bias mitigation approach.

5.3.3 Navigating uncertainty around regulatory expectations, and its consequences. As shown in Section 5.1, while fair lending practitioners understood they needed to test their systems for discrimination and search for LDAs, they also had little concrete guidance on how to do so. This lack of clarity resulted in a significant variety of processes. As E_{11} stated: “*I'm not even sure I've seen two institutions do bias testing and LDA (especially allowable performance loss) exactly the same way; it's kind of wild how different it is.*”

Some participants reported this lack of clarity afforded them room to experiment with new and, in their experience, more effective methods and structures to reduce discriminatory outcomes. As discussed above in Section 5.3.2, participants observed their firms' propensity to stick to established yet inadequate methods, but at the time of our data collection many were beginning to reconsider them (E_1, E_4, T_5, E_9). E_9 shared that “*as we use machine learning models [...] in the industry, everybody is moving away from a manual review [i.e. drop one], because that cannot suffice.*”

Indeed, some participants reported that their firms interpreted the lack of explicit regulatory guidance as an opening to explore different procedural structures to allow some awareness of demographic trends in aggregate while still managing disparate treatment risk. For instance, several participants expressed comfort with hyperparameter-tuning or “pipeline aware” methods [11] to search the design space of models, using demographic measurements only to choose among models after training (E_1, E_4, T_5). E_1 described their firm's LDA search

⁶Following [50], we disagree with this view, but do not address it here as it is out of scope for this work.

process, explaining that they would “*search the entire design space. But obviously, for practical reasons, we’re going to spend only X amount of dollars to search Y number of alternative models, and with the Y number of alternative models that we have done some sort of like semi-exhaustive search...that fulfills our need to show that we could not have done any better.*” E_1 continued, noting that using demographic data as an iterative reward to search the design space seemed not to bring up disparate treatment concerns since it was simply used to measure reward after experimenting with feature swapping, feature engineering, or hyperparameter tuning: “[Legal and compliance] *were happy with that level of arm’s length where we’re not explicitly interfering [...] they were happy to sign off that this is a good methodology for this purpose.*” E_4 , also noted that they had adopted more modern debiasing methods: “*we relied on generative adversarial networks [for debiasing...] that was way more impactful than the old drop-one analysis.*” Had regulators preordained a suite of methods, their guidance may not have been capacious enough to accommodate such exploration, further disincentivizing proactive fairness testing.

At the same time, other participants reported how this same regulatory ambiguity incentivized their firms to do the bare minimum, served as a challenge or blocker to fair lending work, and ended up orienting fair lending teams’ work mainly to document justifications of unaddressed disparate impact that they feel the regulator would accept rather than actually mitigating disparities. Importantly, existing disparate impact doctrine does not require firms to extinguish any disparate impact against a protected group. Instead, the reach of disparate impact doctrine is more limited: it merely requires firms to establish a business justification for the practice that results in a disparate impact. Specifically, under the Equal Credit Opportunity Act (ECOA), a financial institution must show that its practice “meets a legitimate business need.” As a result, E_{11} described that “*many institutions want to have a credible performance of doing the LDA search because they’re mostly concerned with being able to display something that satisfies the regulator.*” But E_{11} noted that “*there is a large gap between ‘we need to demonstrate that we’re trying here’ versus ‘we actually want to do an LDA search that finds better alternatives.’*”

In stark contrast to E_1 ’s firm’s thorough search described above, L_5 ’s firm often didn’t even test models subject to fair lending law because they were confident they could justify typical models as simply common practice: “*For a linear or logistic regression, or even some of the AI techniques [...] if the inputs are basically your credit data, or what is normally collected on an application, as a rule, we don’t feel like there’s there’s real value in doing the statistical testing, in part, because we’re not sure what we would change [...] mostly because the law and the regulatory expectations are actually very unclear.*” Regardless, many participants experienced how fair lending work can end up an exercise in defensive documentation—that is, justifying the legality of an ongoing practice to a regulator, rather than earnestly searching for less discriminatory decision-making systems. In fact, such “justificatory work” appeared to be core to what financial regulators have expected to see in the past, and core to what fair lending programs are tasked to produce. One example that participants described is a firm’s business justification memo, report, or slide deck—an artifact that is often prepared by financial institutions explicitly for a regulatory audience if a model the firm wants to launch has been determined to exhibit a disparate impact. According to T_9 , this document would contain a description of the model, a list of variables, the performance of the model, results of tests conducted, whether or not any viable LDAs were found, and “*if there were, why they weren’t taken, or why they were implemented.*”

Some institutions were described to be very comfortable with justifying disparities using projected profit losses, or by arguing that the practices leading to them were industry standard enough to justify the disparity without even looking for other models (L_5 , E_9). E_9 explained as an example that “*Blacks could have a lower FICO score than Whites. But that’s because the actual default risk that has been observed for the Black class is higher. So that’s one business justification,*” among other factors a firm could use to explain away remaining disparities. Another way participants described their firms defining rules that inhibited the effective prevention of discrimination was to allow for zero performance drop when searching for a less discriminatory alternative. R_1 noted that several companies were “*0 change*”—that is, tolerated no change in performance metrics for an LDA, and justified that

position with their profit interest. They queried whether that justification was truly sufficient, but conceded that such was the act of businesses conducting risk management.

6 Discussion

6.1 More mature practices are still plagued by familiar challenges

While financial institutions' bias mitigation practices appear far more mature than those of entities not covered, financial practitioners still face many of the same problems that have surfaced in entities not directly covered by civil rights law [3, 22, 29, 56, 70, 73]. Though the hurdles are familiar, their trajectory in the financial sector reveal notable differences that offer RAI researchers, as well as practitioners in less mature organizations, a glimpse into a potential future.

One shared problem is limited access to demographic data. While entities not covered by fair lending law typically experience this as simply missing or insufficient demographic data for fairness testing purposes [6, 13], financial institutions have established methods (e.g., BISG) for race data but still lack methods for most other demographics. Sex and other characteristics still protected by law such as national origin and disability are often missing: T_{10} shared that variables such as disability were “*much, much more difficult to collect data on.*” While the use of BISG, despite its recognized limitations, has been accepted by regulators, the existence of such a method at all has created a perverse incentive to avoid collecting self-identified race data. Thus, financial institutions generally lack it (except for mortgage loans, where applicants may disclose this information), even though they have the option to collect it from borrowers [75].

Second, since the decision about whether or not to mitigate disparities rests with the business side of the firm, fair lending practitioners still have to foster company and colleague buy-in informally, similar to RAI practitioners [5, 31, 33, 61, 70], in order to motivate substantive rather than defensive antidiscrimination work. While research has pointed to gaps in regulation to explain corporate inaction on fairness, we observe that even with the specter of regulatory scrutiny to motivate fairness work, participants (L_2 , T_2 , E_4) described needing to leverage appeals to pathos as a way to successfully lobby their organizations to adopt more proactive fairness approaches, posing questions to colleagues and decisionmakers like “*would you really want your grandmother to know that we were doing this?*” (L_2).

Finally, the research-practice gap that challenges ordinary institutions seems particularly acute in the highly regulated financial industry. From a technical perspective, fairness problems in fair lending starkly diverge from those contemplated in the literature: information flow between model builders and testers (first line/second line) about measurement results is significantly more limited than research typically envisions. Testing is typically not done with true demographic data, but imputed proxies, implicating measurement challenges that have received less research attention than other topics. And, practitioners must consider more than the two or four protected groups typically considered in research—fair lending work must be done over a “*2 by 7 matrix*” (E_1) of demographics. This can lead to the “train-crash” problem, where it is combinatorially challenging to simultaneously reduce disparities for all protected groups (E_1 , E_9). While a great deal of theoretical work has addressed this problem [41, 47], practitioner-ready tools have yet to follow.

Beyond these familiar themes, the lack of attention paid in the development of bias mitigation methods to practical tensions between disparate treatment and impact has meant that most algorithmic debiasing methods are seen as too legally risky to adopt in practice, even if thorough technical and legal review would suggest they are not [50]. This tension results in a practical ceiling of fair lending efforts at the model level, leaving fair lending practitioners in a position where using inferior methods for bias mitigation is safest, underscoring the importance of basing research agendas on empirical insights from the field.

6.2 The mixed success of fair lending offers useful lessons for regulatory proposals for AI governance

The history of fair lending offers a cautionary tale for those looking to regulation to solve the many ills of algorithmic decision making. While our empirical findings support the intuition that regulation operationalized through vigorous and proactive enforcement mechanisms can foster desirable practices that firms might not adopt voluntarily, they also highlight that the success of even such potent regulation is not a sure thing. Developing effective regulation is a far more challenging task than the current literature's countless policy proposals might seem to suggest.

To begin, regulation is commonly developed without sufficient attention to administrability—that is, how firms will attempt (or not attempt) to comply with the law in practice [43]. Fully anticipating, *ex ante*, how firms will respond to regulation—as well as the organizational and practical challenges that they will face in seeking to comply with regulation—is genuinely challenging. Even recognizing these challenges, regulatory proposals are often concerningly detached from the realities of compliance. It's not enough for regulation to have “teeth” in theory; effective regulation takes into account how compliance will work in practice. This may involve accounting for natural disincentives to comply fully, with commensurate penalties to shift firms' risk calculus. In other cases, it may require understanding the everyday challenges that practitioners face even when they are sufficiently incentivized to attempt to comply.

Second, regulatory proposals are too rarely informed by evidence demonstrating the relative efficacy of different policy interventions. While there have been recent calls to make AI policymaking more evidence-based (e.g., [14]), these calls tend to focus on evidence about the underlying risks that regulation aims to address rather than about the efficacy of different regulatory designs [8, 24]. Likewise, while there have been recent efforts to evaluate how well regulations targeting algorithmic systems actually work in practice (e.g., [39, 42, 85, 86]), these evaluations rarely engage with questions of what regulatory designs could lead policy interventions to be more effective. Researchers looking to channel their work toward policy recommendations that seek to address the challenges of algorithmic systems should expand their analysis beyond the substance of potential policy requirements and more actively consider the relative successes of different regulatory designs in operationalizing relevant policy goals.

In particular, it is striking how minimally the decades of experience around compliance with and enforcement of fair lending laws have influenced regulatory debates about AI at a structural level.⁷ For example, our findings highlight the unique role that supervisory and enforcement authority have played in successfully fostering fair lending practices—a regulatory design feature that is distinct from other areas of civil rights law and almost completely absent from recent policy proposals for dealing with algorithmic discrimination. Future research could help to avoid such oversight by helping ensure policy interventions and their regulatory designs are directly informed by the lessons learned from empirical studies of compliance, both in historically regulated domains like lending and in the more recent area of responsible AI. To date, the work advancing policy proposals for AI has been largely divorced from the work investigating the problems that AI policy interventions face in practice [43]. While each of these areas of research have reached a point of saturation (the literature on AI policy has ballooned in recent years, while empirical studies of responsible AI in practice have begun to report overlapping findings), there is enormous potential for work that marries the two. More research into the everyday work of enforcement and compliance will go a long way toward ensuring that efforts to craft both substantively and structurally effective regulations are best positioned to succeed.

⁷One potential reason that the decades of fair lending experience has been overlooked is that fair lending supervision is purposefully secretive and confidential. This itself is a trade-off in regulatory design. While such secrecy can empower regulators and regulated entities to share information more candidly, and can even cultivate innovative practices, the secrecy can, absent significant external research, essentially prevent advocates, researchers, and policymakers from understanding the on-the-ground regulatory reality or lock in industry-preferred but less effective practices.

6.3 Discretion is a double-edged sword

Clearly, the regulatory design of fair lending supervision is an area ripe for further exploration by researchers, advocates, and policymakers interested in holding discriminatory AI systems to account. As the findings in this paper suggest, robust supervision and examination can create the conditions for routine, ongoing interventions that fundamentally shape companies' antidiscrimination efforts. While the efforts described by our findings were insufficient in fully combating discrimination, they do point to a potential regulatory path forward.

Nevertheless, regulatory design fundamentally dependent upon the exercise of significant institutional discretion—in particular, discretion as exercised within supervision—can have significant drawbacks. Discretion is not a one-way ratchet toward a static policy goal. In fact, the opposite can be true: discretion is inherently downstream of political ends that may quickly change between administrations. For example, before the CFPB took formal regulatory steps to significantly amend Regulation B and eliminate disparate impact liability [27], and even before the Trump administration promulgated an Executive Order that suggested disparate impact liability “violates our Constitution” [60], the CFPB merely noted in a memo that they “will not engage in redlining or bias assessment supervisions or enforcement based solely on statistical evidence” [27], essentially ending review of potential disparate impact. Such a move highlights how potentially unstable regulatory regimes predicated on the exercise of significant discretion in supervision can be.

7 Limitations

While our sample includes 35 participants, it is not representative of all fair lending practitioners. For example, we did not interview individuals from smaller financial institutions or state regulators. Legal sensitivities and social desirability bias [16] may have also shaped participants' accounts, leading to incomplete descriptions. Further, this study reflects views that have likely evolved since data collection, especially given changes in the federal regulatory environment. Finally, given the lack of prior work, our goal is to offer an empirical account of fair lending practices as they exist on the ground and the factors shaping them, rather than a critical analysis. Future work can and should provide critique.

8 Conclusion

For years, the FAccT community has called for upstream, proactive testing of models that power high-stakes automated decision systems. As our findings show, the regime of fair lending law helped create these very processes, albeit under unfamiliar monikers. Whether readers view these processes as unduly narrow and examples of fairness washing, or exemplars of what ongoing, robust supervision can accomplish, we believe that the multi-decade effort to implement fair lending protections holds important lessons for any effort to combat algorithmic discrimination. Regardless, civil rights law is under attack. Merely reviving previous protections will require concerted effort. In parallel, policymakers are proposing AI-specific interventions to challenge discrimination without the benefit of critical insights from the trenches of fair lending compliance. We believe that both such efforts to strengthen civil rights protections can find value in grappling with the successes and limitations revealed by our results.

9 Generative AI Statement

The authors used generative AI to help format bibtex citations of references selected by the authors. There was no other use of Generative AI assistance in making this manuscript.

References

- [1] 1968. Fair Housing Act, 42 U.S.C. §§ 3601–3619.
- [2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. PMLR, 60–69. Introduces a framework for enforcing fairness constraints via reductions to cost-sensitive classification.
- [3] Muhammad Azeem Akbar, Arif Ali Khan, Sajjad Mahmood, Saima Rafi, and Selina Demi. 2024. Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges. *Software: Practice and Experience* 54, 9 (2024), 1621–1650.
- [4] Pouria Akbarighatar. 2024. Operationalizing responsible AI principles through responsible AI capabilities. *AI and Ethics* (2024), 1–15.
- [5] Sanna J Ali, Angèle Christin, Andrew Smart, and Riitta Katila. 2023. Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 217–226.
- [6] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 249–260.
- [7] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. 2023. “Fairness Toolkits, A Checkbox Culture?” On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 482–495.
- [8] Robert Baldwin, Martin Cave, and Martin Lodge. 2011. *Understanding regulation: theory, strategy, and practice*. Oxford university press.
- [9] Kenneth A Bamberger and Deirdre K Mulligan. 2015. *Privacy on the ground: driving corporate behavior in the United States and Europe*. MIT Press.
- [10] Rebecca Bauer-Kahan. 2025. Automated Decision Systems (Automated Decisions Safety Act). California State Legislature. https://leginfo.ca.gov/faces/billNavClient.xhtml?bill_id=20250260AB1018 A.B. 1018, 2025–2026 Regular Session (Cal.).
- [11] Emily Black et al. 2024. Less Discriminatory Algorithms. *Georgetown Law Journal* 113 (2024), 53–.
- [12] Board of Governors of the Federal Reserve System. 2007. *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit*. Technical Report. Federal Reserve Board. <https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/> Submitted pursuant to section 215 of the Fair and Accurate Credit Transactions Act of 2003.
- [13] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 492–500. doi:10.1145/3351095.3372877
- [14] Rishi Bommasani, Sanjeev Arora, Jennifer Chayes, Yejin Choi, Mariano-Florentino Cuéllar, Li Fei-Fei, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Scott Singer, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, and Dawn Song. 2025. Advancing science- and evidence-based AI policy. *Science* 389, 6759 (31 July 2025), 459–461. doi:10.1126/science.adu8449 Policy Forum.
- [15] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. doi:10.1080/2159676X.2019.1628806 Paper discussing developments and clarifications of reflexive thematic analysis.
- [16] Virginia Braun and Victoria Clarke. 2021. Thematic analysis: A practical guide. (2021).
- [17] Jacques Bughin. 2024. Doing versus saying: responsible AI among large firms. *AI & SOCIETY* (2024), 1–13.
- [18] California State Legislature. 2025. Transparency in Frontier Artificial Intelligence Act, SB 53. State Law.
- [19] Noel Capon. 1976. Credit Scoring Systems: A Critical Analysis. *Journal of Marketing Research* (1976). <https://business.columbia.edu/sites/default/files-efs/pubfiles/690/24.pdf> Columbia Business School Working Paper version.
- [20] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *Comput. Surveys* 56, 7 (2024), 1–38. doi:10.1145/3616865 Comprehensive overview of fairness definitions and methods, organizing pre-processing, in-processing, and post-processing approaches.
- [21] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
- [22] Robert Cinca, Enrico Costanza, and Mirco Musolesi. 2025. Practitioners and Bias in Machine Learning: A Study. *ACM Transactions on Interactive Intelligent Systems* 15, 2 (2025), 1–28.
- [23] Julie E. Cohen, Nina-Simone Edwards, Meg Leta Jones, and Paul Ohm. 2025. *Designing Policymaking Mechanisms for Regulatory Dynamism*. Preliminary Concept Paper. Georgetown University Law Center, Institute for Technology Law & Policy, Reimagining the Governance Stack Project.

- [24] Julie E Cohen, Paul Ohm, Meg Leta Jones, Brenda Dvoskin, and Smitha Krishna Prasad. 2024. Regulatory Monitoring in the Information Economy. *Redesigning the Governance Stack Project* (2024).
- [25] Colorado State Legislature. 2024. Colorado Anti-Discrimination in AI Law / Consumer Protections for Artificial Intelligence (Colorado AI Act), SB 24-205. State Law.
- [26] Consumer Financial Protection Bureau. 2014. Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity. White Paper. https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf
- [27] Consumer Financial Protection Bureau. 2025. Equal Credit Opportunity Act (Regulation B). *Federal Register* 90 (13 Nov. 2025), 50901–50923. Proposed Rule, Document No. 2025-19864, RIN 3170-AB54.
- [28] Peter Conti-Brown and Sean H. Vanatta. 2025. *Private Finance, Public Power: A History of Bank Supervision in America*. Princeton University Press.
- [29] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.
- [30] Wesley Hanwen Deng, Solon Barocas, and Jennifer Wortman Vaughan. 2025. Supporting Industry Computing Researchers in Assessing, Articulating, and Addressing the Potential Negative Societal Impact of Their Work. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–37.
- [31] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [32] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 473–484.
- [33] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating practices and opportunities for cross-functional collaboration around AI fairness in industry practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 705–716.
- [34] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
- [35] Lauren B Edelman. 2020. *Working law: Courts, corporations, and symbolic civil rights*. University of Chicago Press.
- [36] Equal Employment Opportunity Commission. 1983. Procedures for Complaints of Employment Discrimination Filed Against Recipients of Federal Financial Assistance. 29 C.F.R. Part 1691. Originally published 48 FR 3574 (Jan. 25, 1983); last amended 54 FR 32063 (Aug. 4, 1989). Electronic Code of Federal Regulations, current as of Apr. 30, 2026. <https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1691>.
- [37] FinRegLab. 2023. Explainability and Fairness in Machine Learning for Credit Underwriting: Policy Analysis. <https://finreglab.org/research/explainability-fairness-in-machine-learning-for-credit-underwriting-policyanalysis/> Dec. 2023.
- [38] FinRegLab. 2023. Machine Learning Explainability & Fairness: Insights from Consumer Lending. <https://finreglab.org/research/machine-learning-explainability-fairness-insights-from-consumer-lending/> Jul. 2023.
- [39] Marissa Kumar Gerchick, Ro Encarnación, Cole Tanigawa-Lau, Lena Armstrong, Ana Gutiérrez, and Danaé Metaxa. 2025. Auditing the Audits: Lessons for Algorithmic Accountability from Local Law 144’s Bias Audits. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’25)*. Association for Computing Machinery, New York, NY, USA, 29–44. doi:10.1145/3715275.3732004
- [40] Jessica Gonzalez. 2025. New York Artificial Intelligence Consumer Protection Act. New York State Senate. <https://www.nysenate.gov/legislation/bills/2025/S1962> S. 1962, 2025–2026 Regular Sessions (N.Y.).
- [41] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. 2021. Omnipredictors. *arXiv preprint arXiv:2109.05389* (2021).
- [42] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) (FAccT ’24). Association for Computing Machinery, New York, NY, USA, 1107–1120. doi:10.1145/3630106.3658959
- [43] Neel Guha, Christie M. Lawrence, Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, and Daniel E. Ho. 2024. AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. *The George Washington Law Review* 92, 6 (Dec. 2024), 1473–1557. <https://www.gwlr.org/ai-regulation-has-its-own-alignment-problem-the-technical-and-institutional-feasibility-of-disclosure-registration-licensing-and-auditing/>
- [44] Lakshitha Gunasekara, Nicole El-Haber, Swati Nagpal, Harsha Moraliyage, Zafar Issadeen, Milos Manic, and Daswin De Silva. 2025. A Systematic Review of Responsible Artificial Intelligence Principles and Practice. *Applied System Innovation* 8, 4 (2025), 97.

- [45] Patrick Hall, Benjamin Cox, Steven Dickerson, Arjun Ravi Kannan, Raghu Kulkarni, and Nicholas Schmidt. 2021. A United States Fair Lending Perspective on Machine Learning. *Frontiers in Artificial Intelligence* 4 (June 2021), 695301. doi:10.3389/frai.2021.695301 Mini Review article, Section: AI in Finance.
- [46] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, Vol. 29. NeurIPS, 3315–3323. Proposes the “equal opportunity” fairness criterion for predictors in supervised learning.
- [47] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. PMLR, 1939–1948.
- [48] Daniel E Ho and Alice Xiang. 2020. Affirmative algorithms: The legal grounds for fairness as awareness. *U. Chi. L. Rev. Online* (2020), 134.
- [49] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.
- [50] Pauline T. Kim. 2022. Race-Aware Algorithms: Fairness, Nondiscrimination, and Affirmative Action. *California Law Review* 110 (2022), 1539–.
- [51] I. Elizabeth Kumar, Keegan E. Hines, and John P. Dickerson. 2022. Equalizing Credit Opportunity in Algorithms: Aligning Algorithmic Fairness Research with U.S. Fair Lending Regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) (*AIES '22*). Association for Computing Machinery, New York, NY, USA, 357–368. doi:10.1145/3514094.3534154
- [52] Khoa Lam, Benjamin Lange, Borhane Bili-Hamelin, Jovana Davidovic, Shea Brown, and Ali Hasan. 2024. A Framework for Assurance Audits of Algorithmic Systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) (*FAcCT '24*). ACM, New York, NY, USA, 1480–1494. doi:10.1145/3630106.3658957
- [53] Hao-Ping Hank Lee, Lan Gao, Stephanie Yang, Jodi Forlizzi, and Sauvik Das. 2024. “I Don’t Know If We’re Doing Good. I Don’t Know If We’re Doing Bad”: Investigating How Practitioners Scope, Motivate, and Conduct Privacy Work When Developing {AI} Products. In *33rd USENIX Security Symposium (USENIX Security 24)*. 4873–4890.
- [54] Michael Madaio, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1544–1558.
- [55] Michael A Madaio, Jingya Chen, Hanna Wallach, and Jennifer Wortman Vaughan. 2024. Tinker, Tailor, Configure, Customize: The Articulation Work of Contextualizing an AI Fairness Checklist. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–20.
- [56] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
- [57] Edward J. Markey. 2025. Artificial Intelligence Civil Rights Act of 2025. United States Senate. <https://www.congress.gov/bill/119th-congress/senate-bill/3308> S. 3308, 119th Cong., 1st Sess..
- [58] New York City Council. 2021. New York City Local Law 144 of 2021 – Automated Employment Decision Tools (AEDT) Bias Audit Law. Municipal Legislation.
- [59] Ariadne A Nichol, Meghan Halley, Carole Federico, Mildred K Cho, and Pamela L Sankar. 2024. Moral engagement and disengagement in health care AI development. *AJOB Empirical Bioethics* 15, 4 (2024), 291–300.
- [60] Executive Office of the President. 2025. Ending Illegal Discrimination and Restoring Merit-Based Opportunity. Executive Order 14173. <https://www.federalregister.gov/documents/2025/01/31/2025-02097/ending-illegal-discrimination-and-restoring-merit-based-opportunity> 90 Fed. Reg. 8633 (Jan. 31, 2025).
- [61] Will Orr and Jenny L Davis. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society* 23, 5 (2020), 719–735.
- [62] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 39–48.
- [63] Martha A. Poon. 2012. *What Lenders See: A History of the Fair Isaac Scorecard*. Ph.D. Dissertation. University of California. <https://escholarship.org/uc/item/7n1369x2>
- [64] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
- [65] Mark Ryan, Eleni Christodoulou, Josephina Antoniou, and Kalypso Iordanou. 2024. An AI ethics ‘David and Goliath’: value conflicts between large tech companies and their employees. *AI & SOCIETY* 39, 2 (2024), 557–572.
- [66] Seamus Ryan, Camille Nadal, and Gavin Doherty. 2023. Integrating fairness in the software design process: An interview study with hci and ml experts. *IEEE Access* 11 (2023), 29296–29313.

- [67] Malak Sadek and Celine Mougnot. 2024. Challenges in Value-Sensitive AI Design: Insights from AI Practitioner Interviews. *International Journal of Human–Computer Interaction* (2024), 1–18.
- [68] Nithya Sambasivan and Rajesh Veeraraghavan. 2022. The deskilling of domain expertise in AI development. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–14.
- [69] Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. 2023. Don't look at the data! how differential privacy reconfigures the practices of data science. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [70] Daniel S Schiff, Stephanie Kelley, and Javier Camacho Ibáñez. 2024. The emergence of artificial intelligence ethics auditing. *Big Data & Society* 11, 4 (2024), 20539517241299732.
- [71] ANDREW D SELBST and SOLON BAROCAS. 2023. UNFAIR ARTIFICIAL INTELLIGENCE: HOW FTC INTERVENTION CAN OVERCOME THE LIMITATIONS OF DISCRIMINATION LAW. *University of Pennsylvania Law Review* 171, 4 (2023).
- [72] Jessie J Smith, Anas Buhayh, Anushka Kathait, Pradeep Ragothaman, Nicholas Mattei, Robin Burke, and Amy Volda. 2023. The many faces of fairness: Exploring the institutional logics of multistakeholder microlending recommendation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1652–1663.
- [73] Jessie J Smith, Michael Madaio, Robin Burke, and Casey Fiesler. 2025. Pragmatic Fairness: Evaluating ML Fairness Within the Constraints of Industry. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 628–638.
- [74] Texas State Legislature. 2025. Texas Responsible Artificial Intelligence Governance Act, HB 149. State Law.
- [75] United States Congress. 1974. Equal Credit Opportunity Act, 15 U.S.C. §§ 1691–1691f. Federal Statute.
- [76] United States Congress. 1975. Home Mortgage Disclosure Act. Pub. L. No. 94-200, 89 Stat. 1124, codified at 12 U.S.C. §§ 2801–2810.
- [77] United States Congress. 1977. Community Reinvestment Act. Pub. L. No. 95-128, title VIII, 91 Stat. 1147, codified at 12 U.S.C. §§ 2901–2908.
- [78] Ari Ezra Waldman. 2021. *Industry unbound: The inside story of privacy, data, and corporate power*. Cambridge University Press.
- [79] John R. Walter. 1995. The Fair Lending Laws and Their Enforcement. *Federal Reserve Bank of Richmond Economic Quarterly* 81, 4 (1995), 61–77. https://www.richmondfed.org/~media/richmondfedorg/publications/research/economic_quarterly/1995/fall/pdf/walter.pdf
- [80] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In search of the aspiration in data annotation.. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–16.
- [81] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [82] David Gray Widder, Laura Dabbish, James D Herbsleb, and Nikolas Martelaro. 2024. Power and Play: Investigating "License to Critique" in Teams' AI Ethics Discussions. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–23.
- [83] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society* 10, 1 (2023), 20539517231177620.
- [84] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. Limits and possibilities for "Ethical AI" in open source: A study of deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2035–2046.
- [85] Lucas Wright, Roxana Mika Muenster, Briana Vecchione, Tianyao Qu, Pika (Senhuang) Cai, Alan Smith, Comm 2450 Student Investigators, Jacob Metcalf, and J. Nathan Matias. 2024. Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (Rio de Janeiro, Brazil) (FAcCT '24)*. Association for Computing Machinery, New York, NY, USA, 1701–1713. doi:10.1145/3630106.3658998
- [86] Lucas Wright, Roxana Mika Muenster, Briana Vecchione, Tianyao Qu, Pika (Senhuang) Cai, Alan Smith, Comm 2450 Student Investigators, Jacob Metcalf, and J. Nathan Matias. 2024. Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (Rio de Janeiro, Brazil) (FAcCT '24)*. Association for Computing Machinery, New York, NY, USA, 1701–1713. doi:10.1145/3630106.3658998
- [87] Wenbin Zhang. 2024. AI fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine* 45, 3 (2024), 386–395.

A Participant Information

In this section, we include our participant information in Table 1.

B Interview Protocols

The following section provides the protocols used in our semi-structured interviews with policymakers, engineers, and regulators. Note that due to the nature of semi-structured interviews, some questions in the protocol were omitted and follow-up questions were asked instead.

Table 1

Participant	Population	Participant	Population	Participant	Population
P1 (L1)	Lawyer	P15 (E6)	Engineer	P29 (T10)	Third Party
P2 (L2)	Lawyer	P16 (T6)	Third Party	P30 (E8)	Engineer
P3 (T1)	Regulator, Third Party	P17 (T7)	Third Party	P31 (E9)	Engineer
P4 (T2)	Third Party	P18 (R1)	Regulator, Engineer	P32 (R7)	Regulator
P5 (L3)	Lawyer	P19 (R2)	Regulator	P33 (L7)	Lawyer
P6 (E1)	Engineer	P20 (R3)	Regulator	P34 (E10)	Engineer
P7 (E2)	Engineer	P21 (R4)	Regulator	P35 (R8)	Regulator
P8 (E3)	Engineer	P22 (T8)	Third Party		
P9 (L4)	Lawyer	P23 (E11)	Regulator, Engineer		
P10 (T3)	Third Party	P24 (L5)	Lawyer		
P11 (T4)	Third Party	P25 (L6)	Lawyer		
P12 (E4)	Engineer	P26 (T9)	Third Party, Lawyer		
P13 (T5)	Third Party	P27 (R6)	Regulator		
P14 (E5)	Engineer	P28 (E7)	Engineer		

Table 1. Participants' professional backgrounds. Following prior work on responsible AI practices in industry [33, 55, 82], we omitted demographic details and abstracted certain information about participants' employers and roles to avoid inadvertently identifying individuals working at the forefront of this high-stakes domain.

B.1 Policymakers

Background, process, and organizational structure

In this interview, we're interested in the process by which models are trained and evaluated, especially with regards to fairness, and whether and how legal and compliance interacts with these processes. We appreciate that you're not able to share specific legal advice or information protected by legal privilege and just to emphasize we aren't looking for you to share that information, we are interested in more general concepts ; feel free to phrase your answers in the hypothetical or in general in order to answer to whatever degree of detail you are able to.

Q1 Could you tell us briefly about what your role is at your current organization and what brought you to this type of work?

Q2 We are talking with people at companies across different contexts, for example financial services, online platforms, housing, and more. What application space is your team or company working on?

We're interested in learning about different processes different firms have related to algorithmic fairness or discrimination. First, we'll be asking about your firm's approach.

Q1 Is there a structured process by which teams at your organization are performing work related to algorithmic fairness or discrimination of any kind?

- (1) Does your organization perform any tests for algorithmic discrimination, for example testing for disparities in model behavior across demographic groups? Why or why not?
 - (a) If so, talk us through that testing process. What triggers the testing process to start?
 - (b) When are tests performed, and what tests are done?
 - (c) What, if any, is your firm's policy in the case that a disparity is discovered?
 - (d) When, if ever, would your firm look for less discriminatory models to replace a potentially discriminatory model? How would this search take place?

Q2 Does your organization have defined policies, practices, or processes for decision making when conducting work on AI fairness or discrimination? For example, does your organization have written policies and procedures about how to evaluate and mitigate disparities or search for fairer models?

- (1) Do you know how your organization came to these policies and processes?
- (2) Does your organization have any norms or explicit policies around documenting algorithmic fairness efforts?

For the next few questions, we're curious about your firm's organizational structure around fairness testing.

Q1 Is there a fairness/anti-discrimination/fair lending team? If so, how is that team structured? What is the relationship between this team and other teams, including model developers?

- (1) If so, how is that team structured?
 - (a) What is the relationship between this team and other teams, including model developers?
 - (b) What role does this team play: are they a gate, guide, or partner?
 - (c) Who is responsible for implementing the team's recommendations?
 - (d) Whom does that team report to?
 - (e) Do you know why the work was organized in this way? For example, were there any legal considerations?
- (2) If not, where does fairness work occur in the organization?
 - (a) What role do those who do fairness work play– guides, gates, or partners?
 - (b) Who do the fairness practitioners report to?
- (3) How are you personally involved in this work? Where do you come in in this structure?

Q2 Who makes key decisions about this work? For instance, setting priorities, methods, policies, etc.

- (1) Who is responsible for making decisions about whether or not a given model is deployed?
- (2) Are you involved in any decisions for what measurements or mitigations are necessary, or which models are ultimately appropriate to deploy?
 - (a) If yes: What factors tend to inform the recommendations you would generally make in these sorts of conversations?
 - (b) If no: Can you give me your sense of what goes into these decisions?

Q3 Given that organizational structure, how do these teams (either fairness/fair lending/ or individuals who do fairness work outside of a team) interact with other teams at your organization – in particular ML/AI developers – to conduct fairness work?

- (1) What is the relationship between these teams like?
- (2) Are there any things you've observed to be particularly challenging when less-technical stakeholders communicate with technical colleagues as they do this work?
- (3) Are there any other challenges you've experienced that are relevant to fairness work at your organization?

Q4 Are there any internal policies that seem to constrain your organization's ability to test and mitigate algorithmic disparities?

- (1) How do you think your organization's process as it stands could be improved to fulfill its higher order goals of preventing discrimination?
- (2) What do you think would need to change in order to accomplish these goals?

Laws that underwrite the process (DI) and interpretations/ regulatory expectations

As we mentioned, we're interested in how models are evaluated and trained, especially with regards to fairness, and whether and how legal and compliance interacts with these processes. For the next few questions, we're interested in how laws, regulations, and policies shape these efforts.

Q1 Disparate impact has informed a huge range of algorithmic fairness/fair lending/anti-discrimination work – how do you think that disparate impact doctrine relates to your work? Does it inform the work your firm does around fairness testing?

- (1) If so, tell us a bit more about what you believe it requires your organization to do, or not do?
 - (a) How does the doctrine relate to the process by which your firm tests for algorithmic discrimination?
 - (i) How does the doctrine relate to searches for fairer models?
 - (ii) Do different models or model types fall under or outside of the DI framework at your organization? What products or models don't? What makes these products or models different from one another? For example, is there an explicit classification of low risk/high risk products? How are models classified into each?
 - (b) How does your team think about the second step of the disparate impact test – the business justification of a model?
- (2) If not, do you have a sense of why this doctrine is not seen to directly apply to your business?
- (3) <If they have a fairness testing/mitigation process> What motivated your company to implement discrimination testing (and mitigation), given the lack of a legal requirement?
 - (a) In absence of a legal framework, are there any principles your firm uses to guide the testing and mitigation of algorithmic bias? For example, are there any particular factors your organization considers with regard to the defensibility of launching a product that may have a disparate impact or launching a particular version of a model over a different version?

Q2 In your opinion, are there scenarios where it would be clear that a team like yours, or an organization like yours, would be required to search for an LDA?

- (1) If yes, does that happen today?
- (2) If no, can you imagine any?
- (3) In thinking about LDAs that serve the same business purpose, what does “serve the same business purpose” mean to you? Can you talk about what that concept means to you in your work?

Q3 Are there laws, regulations or guidance that seem to constrain your organization's ability to test models for potential performance differences across demographic groups?

- (1) If you could make changes to such laws, what would they be and why?

Q4 Do you believe that regulatory agencies (such as, e.g. the CFPB/EEOC/FHA) have certain expectations of how your business performs fairness tests and/or bias mitigation? Why or why not?

- (1) What do you perceive those expectations to be, if so?
- (2) For the relevant regulatory agency in your domain (e.g. CFPB, HUD, FTC), how do you keep up to date on any guidance whether formal or informal?

Q5 Do you think your organization, or organizations like yours, would benefit from more regulatory clarity or guidance on how to search for LDAs? What would that look like, hypothetically?

Cost and other constraints

Q1 At organizations like yours, to what extent are costs actively considered when determining what models should be tested?

- (1) How about the costs involved in searching for fairer models (LDAs), or what model to choose after searching?
- (2) What are some of the costs that you're aware of that play a role in decision-making?

Q2 Are some more important than others?

- (1) How does your organization navigate these costs in the context of compliance?

(2) Is this consideration explicit or subtextual?

Q3 Are there other factors beyond cost that are considered when determining which models to pursue or to launch?

- (1) Can you share about what sort of factors are considered, and to what extent?
- (2) Do any particular factors tend to trump others?

Wrap up

Q1 Is there anything you wanted to say that you didn't get the chance to?

Q2 Is there anyone else you could connect us with that might be a good fit for our interview?

B.2 Engineers

Background and overall process

Q1 Could you describe what application space your team is working on?

Q2 Does your organization have written policies and procedures about how to evaluate your models? How about how to mitigate disparities or search for fairer models?

- (1) Is there a designated team who's responsible for this? Are you a part of this team?
 - (a) What is the relationship between this team and your team?
 - (b) Who do you report to?
 - (c) And who does the fairness team report to?
- (2) Do you have legal and compliance teams? How big are the teams? Who do they report to?
- (3) Do you have an inventory of all your models?

Q3 Imagine you're deploying a new model in your team <insert appropriate concrete example>. Can you briefly take me through the process of testing, reviewing, and deploying that new model?

- (1) What would be the decision points about specifically what model versions to launch?
- (2) Do you do any testing of your model or model versions across different populations?
 - (a) What are the different groups? Do they have any relationship to demographic groups?
 - (i) Which demographic groups and why?
 - (ii) How do you measure demographic groups?
 - (b) Under what conditions do you do this testing?
 - (c) How broad is the testing?
 - (i) How much time does it take and how many people are involved?
 - (ii) Do you have a sense of how much it costs the organization?
- (3) Does your model building process involve proactively testing for differences across demographic groups, or does it do so in an ad-hoc fashion? Or, does it just happen at the end?
 - (a) Does your team perform these tests or is there some other team that does?

Q4 <If they test for fairness problems> How did you learn what you should be looking out for when performing fairness/bias assessments?

- (1) What are the sources of authority on how analyses are done?
- (2) Do the recommended tools to search for bias change over time?
- (3) Do you keep up with scholarship on where to look for bias?

Metrics

Thanks for providing us with the overall process of your model testing! Now we want to dive a bit deeper into the metrics.

Q1 Are there specific metrics your org/team considers in building models?

- (1) Given the metrics that your team/org considers when determining the quality of a model, what do you think it means for two different model versions to be interchangeable or similarly good?
- (2) How similar do models generally need to be for you to have flexibility to choose between model versions, for example to preference models with more equal behavior across groups
- (3) When you're deciding to deploy a model, what are the most important metrics to perform well on?
 - (a) Are there ranges of values of those metrics, or a threshold?
 - (b) Do you consider any counter-metrics?
 - (c) Are there any particular metrics which are considered to be business-critical?
 - (d) Do you ever consider financial factors as a counter metric? If so, what factors?

Model comparison

Now we've talked about the metrics, let's discuss more about the concrete model comparison based on these metrics

Q1 To what extent does your organization have infrastructure to systematically compare different model versions?

Q2 Do you have any decision-making process in place for how to choose between model versions?

- (1) In discussions that are working through decision-making processes, are there ever conversations that are weighing general model performance versus performance across groups?
- (2) Are there justifications for differences across groups discussed as a part of that process?
- (3) Who ultimately decides if the justifications are reasonable?

Q3 If your team finds disparities as a result of the testing, what are the next steps?

Addressing more fair models (technical) *So far, we've been talking about testing. Now we want to shift the conversation to how you and your organization go about addressing your findings or proactively searching for alternative models or approaches. <If they offer no words for this themselves> We're generally interested in the practice of mitigating disparities found in models to be deployed, or finding equally performant models with fewer disparities across predefined demographic groups. Let's call those more fair models, or MFMs. Imagine looking for MFMs in your team.*

Q1 When looking for MFMs, do you have a baseline model that you try to mitigate or fix? Or, do you generate and compare a wide range of models?

- (1) Can you describe what this process looks like in practice (for baseline or generating a large set of models)? Are there any standard methods you do (or don't) use to mitigate disparities / generate more fair models?
- (2) <If generating a large set of models> Do you try looking for MFMs in each or multiple model building steps, or do you only look into it at the end once the model is made?
 - (a) If so, how?
 - (b) What technical infrastructure do you have to search for MFMs at each of these steps?
 - (c) Do you explore various feature sets?
 - (d) Do you perform hyperparameter tuning with fairness as a metric?

Addressing more fair models (organizational) *Ok, now that I have a better understanding of the technical details on choosing more MFM, let's talk more about the organizational components around this topic.*

Q1 If there is a team that's empowered to search for MFMs, do they have a set of parameters within which they can operate, like a performance or financial budget?

- (1) Are there any reasons that an MFM might not be deployed if found?

Q2 Are there times when you're required to search for an MFM? Or are there times when you feel constrained from searching for one?

- (1) Have you or your colleagues experienced any barriers to searching for MFMs?

- (2) <If there is a legal/compliance/fairness team that is not them>: What is your relationship like with <legal/compliance/fairness team>? Do you ever interact with them directly? What are some tough points?

Q3 <URL> Here are some costs organizations might consider when searching for new models. Are these a part of your day to day work, and are there budgets for any of them? Or, is there a different cost metric your organization uses?

Q4 Do you have any process in place to document the policies, practices, or decisions you make when searching for MFMs?

Q5 Who, in your mind, are the key decision makers who decide whether a model will go into production?

- (1) Does this set of stakeholders change under a different set of circumstances– for example when deciding whether to deploy more fair models?

Introducing Less Discriminatory Alternatives *Great! Now this is the last part of today's interview, and we want to ask you if you are familiar with the concept of Less Discriminatory Alternatives.*

Q1 Are you familiar with the concept of less discriminatory alternatives?

- (1) Does your organization frame its fairness work around the concept of less discriminatory alternatives?
(2) Does what we've talked about so far (with MFMs) relate to how your organization handles a search for LDAs?
(3) If your org does not think about things this way, have there been discussions about whether to adopt such an approach? What's your hypothesis as to why the organization has not yet used this approach?
(a) If organization does, how many times have you seen an LDA replace a baseline model?
(4) Does this process seem to reflect a robust and thorough approach to searching for and mitigating disparities, or does it seem like a narrow and compliance-based approach?

Wrap up

Q1 Is there anything you wanted to say that you didn't get the chance to?

Q2 Is there anyone else you could connect us with that might be a good fit for our interview?

B.3 Regulators

Background and overall process

Thanks again for doing the study. We appreciate that you are speaking in your personal capacity, nothing reflects the views of the regulator or agencies that you have worked for or currently work for. And just to emphasize: we aren't looking for you to discuss any specific investigations or names during our conversation, our questions are aimed at more general concepts. For context, we are doing interviews with people who occupy different roles working with algorithmic fairness, such as engineers who design these systems, product managers who help deploy them, policy and compliance managers who help set policy and practice at companies, and regulators such as yourself. We'd like to start with some basic background questions.

Q1 Could you tell us briefly how you came to work at <agency> and what you do/did?

Q2 Given that background, do you have any experience in your capacity at your current agency or previous agencies with algorithmic discrimination?

- (1) If yes, please elaborate more, perhaps telling us how you've approached the issue.
(2) If no, what about experience with discrimination at <agency>?
(3) Could you at a high-level talk about your agency's overall approach on discrimination in algorithmic systems?

Experience, expectations, and merging the two [*Thanks for that background. We'd like to learn more about your experiences with covered entities in relation to LDAs. For the next few questions, we're really focused on what you have seen companies do in practice.* **Q1** Practically speaking, when it comes to testing algorithmic systems for discrimination, what have you seen regulated firms do? Can you describe the efforts you have seen in practice?

- (1) What kinds of disparities were tested for?
- (2) How have you seen or observed firms measure disparities?
- (3) What are they doing, if anything, if they identify disparities?
 - (a) Are they explicitly searching for less discriminatory alternatives?
 - (b) How, exactly, have they performed these searches?
- (4) How commonly are these performed across firms? How frequently does each firm perform these tests?

Q2 What do you think accounts for the differences in firms' practices that you have observed?

- (1) In your experience, what influences firms' decisions around conducting tests for discrimination in their algorithmic systems?
 - (a) Do firms have different interpretations about what they are legally obligated to do?
 - (b) If interpretations seem to differ, why do you think they differ?
- (2) In your experience, what limits or complicates firms' efforts to test for discrimination in their algorithmic systems?
 - (a) Were there any cases where firms did not do some element of it but ended up having some kind of explanation that the regulator found to be persuasive or reasonable?

Q3 In your experience, what do firms do when they identify a disparate impact?

- (1) In particular, how do they assess whether there is a business necessity justification for the disparate impact?
- (2) Do they attempt to find a less discriminatory alternative?
 - (a) When do firms feel that it is necessary to search for less discriminatory alternatives?
 - (b) Do they perform searches for any model or only specific kinds of models?

Q4 When firms search for less discriminatory models, what have you observed them do in practice?

- (1) Where does your sense come from? Conversations? Regular engagement?

Q5 In your experience, what limited or complicated firms' search for a less discriminatory alternative model?

- (1) To what extent are these challenges similar to or distinct from those that may limit or complicate efforts to test for discrimination in algorithmic systems?

For the next few questions, we're really focused on what you expect companies or want companies to be doing.

Q1 When it comes to testing algorithmic systems for discrimination, what were your agency's expectations of covered entities?

- (1) What's the reason for those expectations being the expectations? Do you know why those were the expectations?
- (2) How were those expectations communicated to covered entities? Were they?

Q2 Imagine a firm has identified a potential disparate impact — what would you expect them to do to justify their business practice?

- (1) What would you expect them to document or provide you?

Q3 When it comes to searching for less discriminatory models, what were your agency's expectations of covered entities?

- (1) Was the expectation that LDAs need to be equally effective in performance to be considered viable alternatives?
 - (a) If yes, what informs your understanding?

(b) If not, what are the requirements on LDA performance?

Q4 What guides your agency's <enforcement/supervision/research> on algorithmic discrimination?

- (1) What was your agency's investigatory capacity like?
- (2) What does it look like to do target selection at your agency?
 - (a) Were there any targets or potential targets of investigation that did not occur due to perceived lack of legal viability?
- (3) To what extent do you have the capacity to conduct fairness testing?
 - (a) Are there any practical barriers to leveraging that capacity?
 - (b) If there is not enough capacity, what happens?
- (4) What other "tools" are available to your agency besides enforcement or supervision? Such as rulemaking, publishing policy statements, circulars, etc.

Q5 What sort of documentation would you expect a firm to create and maintain about their efforts and decisions? What do you expect companies to document and justify the business necessity of their algorithmic systems?

- (1) If yes, could you describe what you'd ideally expect to see in a search?
- (2) If none, could you describe what you'd expect to see in a search?

Q6 Are there any circumstances that you recall where there was a disagreement between a regulator and a covered entity about whether a viable LDA exists? What was the source of the disagreement and how did it resolve?

For the next few questions, we're focused on what the agency has been doing to get companies to actually meet those expectations.

Q1 In the first part of our conversation, we talked about things you saw private firms do. In the second part of our conversation, we talked about what the regulatory expectations were. In your opinion, what is the agency doing/what did the agency do to ensure that companies meet those expectations?

Q2 In particular, in your experience, what do you think has hindered regulatory efforts to clarify that companies should be regularly testing their models for DI and searching for LDAs?

- (1) Specifically, do you think the interplay between disparate treatment and impact has done so?
- (2) To what extent have organizations in your jurisdictions noted or sought guidance around challenges related to the use of demographic data for measurement or mitigation of discrimination?

Q3 Do you believe that further clarification/standardization of requirements would motivate covered entities to be more proactive in searching for LDAs?

Q4 In your experience, is there certain guidance or other "papers" that companies have asked you to provide to clarify expectations?

- (1) When writing guidance, what factors go into the specificity with which these guidances are written, especially with regards to more technical concepts like bias testing and LDAs?

Q5 To what extent, in your experience, do firms react to "sub-regulatory" documents, e.g., blogposts or public statements? <Priming examples below>

- (1) Performance cost: accuracy changes
- (2) Startup cost: having necessary infrastructure/ people in place
- (3) Search cost: training runs
- (4) Deployment cost: changing the deployment to the new model, and the cost of running that model; retraining people to use the new model
- (5) Human cost: cost of making the decision to change the model, and cost of "people-hours" (time spent, salary) on conducting the search

Code/Sub-code	Code/Sub-code
Policies (rules)	Documentation
Trigger for assessment Assessment policy Trigger for mitigation Mitigation policy Deployment policy	Documentation of policy Assessment report Mitigation report Audience and access Attorney client privilege
Procedures	Regulation from Regulators' Point of View
Access to and availability of data Model development Model validation Demographic data Bias testing Bias mitigation Trade-offs	Expectations Regulatory clarity Promote compliance Reasonableness assessment Legal basis Enforcement Willingness to tolerate experimentation
Processes (Practice/standards)	Organizational Dynamics
Feature review process Disparity review process Business necessity Mitigation strategy Determining what counts as a viable alternative Model deployment	Decision-making authority Risk tolerance Division of labor Team structure and expertise Fostering buy-in Market position Cross-functional interactions/collaborations/conflicts
Regulation from Entities' Point of View	Financial Cost
What laws are thought to apply Perceived expectations of regulators, both now and in the future Regulatory clarity Legal risk Political landscape	

Table 2. The codebook resulting from our reflexive thematic analysis, organized by code in bold, followed by its sub-codes.

Closing thoughts

Q1 How do you or your agency keep up to date on the state of the art regarding algorithmic testing, LDAs, and civil rights law?

Wrap up

Q1 Is there anything you wanted to say that you didn't get the chance to?

Q2 Is there anyone else you could connect us with that might be a good fit for our interview?

C Codebook

Our codebook can be found in Table 2.