Auditing the Audits: Lessons for Algorithmic Accountability from Local Law 144's Bias Audits

Marissa Kumar Gerchick
American Civil Liberties Union
New York, NY, USA
mgerchick@aclu.org

Ro Encarnación
University of Pennsylvania
Philadelphia, PA, USA
rone@seas.upenn.edu

Cole Tanigawa-Lau
Stanford University
Political Science
Stanford, CA, USA
coletl@stanford.edu

Lena Armstrong
Harvard University
Cambridge, MA, USA
larmstrong@g.harvard.edu

Ana Gutiérrez
American Civil Liberties Union
San Francisco, CA, USA
agutierrez@aclu.org

Danaé Metaxa
University of Pennsylvania
Philadelphia, PA, USA
metaxa@seas.upenn.edu

Abstract

In this work, we "audit the audits," analyzing the documents produced pursuant to one of the United States' first enacted laws regulating the use of artificial intelligence in employment: New York City's Local Law 144. This law requires employers and employment agencies using certain types of automated tools to publish "bias audits" with statistics about how different sex and racial groups fare in the hiring process when the tools are used. We collect and conduct a comprehensive analysis of all Local Law 144 bias audits (N=116) made publicly available to our knowledge from the law taking effect in July 2023 until early November 2024, and describe the extensive challenges we faced in identifying, archiving, extracting information from, and ultimately analyzing these bias audits. We identify several ways that bias audits produced in accordance with Local Law 144 are incomplete evaluations of algorithmic bias, despite news coverage and characterizations by employers and vendors suggesting otherwise. We show that Local Law 144 bias audits are significantly hampered by several issues, including missing demographic data, opaque data aggregation, problematic uses of "test data," and reliance on metrics that do not represent how automated hiring tools are used in practice. We analyze the reported results in Local Law 144 bias audits alongside the four-fifths rule often used as a measure for assessing adverse impact in employment contexts. Most audits do not report results that would not suggest violations of the four-fifths rule. Crucially, however, we show that these tools could often be in violation of the four-fifths rule when considering potential impacts of missing demographic data. We offer ten practical recommendations to strengthen future legislative efforts that mandate algorithm auditing in hiring and other areas, and contribute an open dataset and codebase for extracting and combining bias audit results to support future auditing efforts.

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732004

CCS Concepts

• **Social and professional topics → Governmental regulations**; • **Computing methodologies → Artificial intelligence**.

Keywords

Local Law 144, algorithm audit, bias audit, automated employment decision tool, four-fifths rule

ACM Reference Format:
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 *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715275.3732004

1 Introduction

Regulators around the world have recently started proposing, advancing, and enacting myriad policies focused on automated systems that use artificial intelligence (AI) and machine learning (ML) [75, 98]. One such effort is New York City's Local Law 144 (LL144), which took effect in July 2023 and focuses on the use of automated tools in employers' hiring and promotion processes [92]. Employers today rely on various kinds of automated systems [16, 25, 103], and prior work has shown that these tools can enable or exacerbate discrimination based on race [43, 78, 115], disability [27, 115], sex [36, 65, 113], and other characteristics, as well as their intersections. LL144 was motivated by the use of such tools, and in the words of the enforcing agency, the NYC Department of Consumer and Worker Protection (DCWP), "prohibits employers and employment agencies from using an automated employment decision tool [(AEDT)] unless the tool has been subject to a bias audit within one year of the use of the tool, information about the bias audit is publicly available, and certain notices have been provided to employees or job candidates" [92].

Shortly after LL144 took effect in the summer of 2023, news outlets optimistically characterized the new law as requiring that companies prove that their AI hiring software "isn't sexist or racist" [33] and "is free from racial and gender bias" [86]. More recent coverage, however, paints a less rosy picture, arguing that Local Law 144 has "largely failed" [112] and that "everyone ignores" the law [99]. In light of these shifting characterizations, interrogating the bias audits produced pursuant to Local Law 144 is necessary for understanding the law's implications, especially given that advocates, journalists, researchers, and even auditors who have conducted LL144 audits themselves have argued these audits are not holistic evaluations of automated hiring tools [42, 59, 87, 123].

In this work, we audit the audits, contributing a novel mixed-methods study of all bias audits made publicly available (to our knowledge) as of early November 2024, collecting and analyzing 44 bias audit reports covering 116 bias audits of automated tools (see Table 1 for more detail). We found audits for approximately 2% of the Fortune 500 companies, despite an industry report finding more than 98% of Fortune 500 companies use applicant tracking software, which often leverages automated systems or artificial intelligence to process candidate materials [100]. We use quantitative and qualitative methods to critically analyze these public documents for their strengths and weaknesses. We identify areas for improvement in each stage of the auditing and reporting process and offer recommendations for strengthening the structure, robustness, usability, and accessibility of legislatively mandated algorithm audits and the policies mandating them (summarized in Figure 1). We also contribute a dataset that includes the results of all the bias audits in our sample, with accompanying code and documentation, available at https://github.com/aclu-national/auditing-the-audits-facct-2025. We ultimately argue that these bias audits produced from Local Law 144 are inadequate as evaluations of potential bias and summarize the lessons they provide about the potential and pitfalls of audit-mandating legislation.

2 Background and Related Work

We first situate our analysis within related literature on auditing algorithmic tools used in hiring, providing context for the key requirements of LL144's bias audits and evaluations to date of the law's impacts.

2.1 Auditing for Bias in Algorithmic Hiring

With the widespread use of automated systems in hiring—including systems leveraging AI—a body of evidence has emerged about the potential for these algorithms to enable or exacerbate discrimination [14, 19, 43, 80, 107, 108, 115]. There is also a long history of discrimination in hiring offline, [15, 23, 25, 35] and systemic discrimination in hiring persists to this day [76, 77]. Proposals to address algorithmic discrimination in hiring today often leverage established guidelines related to the use of hiring tools [37, 38, 64]. One such guideline used in the U.S. by the Equal Employment Opportunity Commission (EEOC), the federal agency tasked with enforcing many workplace civil rights laws, is known as the four-fifths rule. It states that "a selection rate for any race, sex, or ethnic group which is less than four-fifths of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact" [96]. The four-fifths rule has widely shaped algorithm auditing standards in hiring, even though the rule is not the legal standard for discrimination (as the EEOC has stated, the rule is "merely a rule of thumb" [37]) and the shortcomings of the heuristic and its application have also been debated [61, 91, 101, 120]. New standards have also been proposed to address algorithmic bias in hiring [40], including algorithm audits and impact assessments [13, 17, 89, 104, 122], drawing on the growing body of work focused on algorithm auditing more generally [117]. Several legislative proposals, including LL144, mandate that automated systems used in hiring be regularly assessed for disparate impacts [31, 105, 127].

However, concerns have emerged about the efficacy of algorithm audits to combat hiring bias, including the standards of LL144 [42, 87, 88]. These include the potential for "audit washing" [46], risks of corporate capture of auditors [34, 121, 124] and the possibility of "discrimination-hacking", akin to *p*-hacking in statistical testing, where auditors specifically select data or models that have more desirable metrics to present a misleading picture of potential biases [24]. Several scholars have argued that focusing on quantitative outcome metrics (like selection rates or compliance with the four-fifths rule) or auditing particular automated hiring tools in isolation can be misleading [18, 120]. The hiring process often involves interacting with human decision-makers and multiple automated tools that are interconnected [110, 111]—not to mention that some job seekers can circumvent automated hiring processes through informal channels [22, 30]. As a result, audits of automated tools in isolation are incomplete [25].

2.2 Local Law 144 Bias Audit Requirements

To contextualize our analysis, we provide an overview of the key requirements of LL144 bias audits. LL144 bias audits are required to report statistics about how applicants in different sex and race/ethnicity groups fare in the hiring or promotion process [93, 94].[1] The law distinguishes tools used to advance applicants in a hiring or promotion process (*selection tools*) from tools used to score candidates (*scoring tools*). For selection tools, the audits must include the rate at which individuals in a race and/or sex category are automatically selected to move forward or assigned a classification by an AEDT. For scoring tools, the audits must include the rate at which individuals in a race and/or sex category received a score above the sample's median score, when those applicants' scores were assigned by an AEDT. For both selection and scoring tools, the *impact ratio* must also be reported in the audit, defined as the selection or scoring rate for each race and/or sex category divided by the selection or scoring rate of the most selected race and/or sex category [94]. As an illustrative example, if 100 male applicants and 100 female applicants apply for a role where an AEDT is used for selection, and 50 male applicants and 40 female applicants advance, the selection rate is 0.5 for male applicants and 0.4 for female applicants, and the impact ratio is $0.5/0.5 = 1$ for male applicants and $0.4/0.5 = 0.8$ for female applicants. The four-fifths rule, discussed in Section 2.1, can be thought of as a heuristic applied to the impact ratio; if the impact ratio is below 0.8, the four-fifths rule is violated. Importantly, LL144 requires that *bias audits* be conducted, but only requires that *bias audit summaries* be published on employers' websites, and LL144 does not require that selection/scoring rates or impact ratios adhere to any particular threshold [94].

[1] LL144 requires that the relevant statistics be reported for "each race/ethnicity and sex category that is required to be reported on to the U.S. Equal Employment Opportunity Commission ("EEOC") pursuant to the EEO Component 1 report" [94]. Federal courts have recognized that sex discrimination includes discrimination based on sex-based characteristics, including gender identity [114].

STAGE	FINDINGS	RECOMMENDATIONS
SEARCHING	• No standard URL schema • No central repository of audits • No centralized list of deploying entities	1. Central repository of audits maintained by enforcement agency and requirement for employers to post audits using standard URL format to improve audit findability
ARCHIVING	• Prior year audits not preserved when new audits released • Audits posted and later removed	2. Requirements for historical preservation, including submitting new audits to enforcement agency, posting at standard URL, and not removing old audits
EXTRACTING	• Inconsisent audit report and table formatting • Missing or inconsistently formatted audit metadata • Multiple tools listed under one audit • Use of inconsistent race/ethnicity categories • Potentially inaccurate selection rate/impact ratios	3. Standardized audit and table format, posting audits in a machine-readable, accessible format with structured information (including tool and version identifier, date of first use, and time frame for when audit data was collected or created)
ANALYZING	• Focus on race and sex only (not disability, age, etc.) • Missing applicant demographic data • Intra-audit and inter-audit silent duplicates • Ambiguity around data pooling and test data use • Impact ratio interpretations and inconsistencies • Challenges with temporal analysis • Misleading audit characterizations/conclusions	4. Assessing bias for broader set of protected groups 5. Greater auditor access for more holistic auditing 6. Expanded analysis of missing demographic data, including results for missing groups and calculating uncertainty bounds 7. Transparency around data aggregation and using more representative bias metrics 8. Guidelines on how audits can be characterized by auditors
SYNTHESIZING	• Lack of enforcement and required action based on audit findings	9. Transparency into deploying entity governance 10. Actionable requirements and robust enforcement

Figure 1: A summary of problems we encountered in searching for, archiving, extracting information from, and analyzing Local Law 144 bias audits (discussed in Sections 3 and 4 with associated recommendations in Section 5). We encountered issues at each stage of the process, with many challenges compounding, contributing to shortcomings in the audits' effectiveness.

2.3 Evaluations of Local Law 144

Several analyses of LL144 have been published to date, focusing on issues with the law that emerged before, during, and after the law's passage. Before LL144 took effect, various researchers, civil rights advocates, auditors, and industry actors weighed in on the draft proposal, shaping its ultimate form. Groves et al. [59]'s qualitative analysis of public comments on the rule-making process for LL144 revealed tensions between employers, independent auditors, civil society groups, and New York City DCWP (the enforcing agency for LL144) over defining the scope of the law and requirements for independent auditing. For instance, they highlight how industry lobbying shaped the law's definition of AEDTs and its lack of focus on protected characteristics (besides race and sex) where discrimination in hiring and employment has long been a concern, such as disability, pregnancy, and age. Filippi et al. [39] discuss the proposed and final metrics required in bias audit reports, proposing alternatives that, they argue, are more robust to variations in data distributions and better account for thresholds or cutoff scores.

Since the law was enacted, researchers have highlighted challenges in interpreting and complying with the law. Proposals for refining LL144's auditing requirements have begun to emerge, including from entities commonly commissioned to produce LL144 bias audits [81]. Groves et al. [59] also examined audit providers' experiences with the law, who described challenges with accessing data needed for audits and expressed concern that the law's significant room for interpretation may give bad actors an opportunity to engage in cherry-picking or malicious compliance tactics. Relatedly, Cen and Alur [29] highlight how LL144 is emblematic of broader challenges around data access and interpretation of algorithm audit results. Wright et al. [123] report qualitative findings on the experience of systematically searching for LL144 audit reports and transparency notices; at the time of the study, the authors had identified 13 unique audit reports and 13 transparency notices, even though automated hiring technologies are widely used by employers [103]. The authors note that, because employers have so much latitude over deciding whether their automated hiring tools are AEDTs under the law, the relative lack of audits that have been publicly identified so far cannot itself be considered indicative of non-compliance, a phenomenon they instead dub "null compliance." They also highlight that many employers placed these audits and notices in different sections of their websites, making them challenging for job seekers to locate. Across the 13 unique audit reports Wright et al. [123] collected, there were 386 impact ratio measurements; 96% of the impact ratios exceeded 0.8 (implying consistency with the four-fifths rule the vast majority of the time), and 20% of the impact ratios had missing information, primarily for marginalized Indigenous groups.

To this landscape, we contribute a novel analysis of all 116 bias audits published to our knowledge from July 2023 (when the law was enacted) through November 2024. Our findings not only underscore the themes of existing work on LL144—including that the bias audits are inaccessible, uninterpretable, and characterized by missingness—but also surface novel insights with practical value for future legislative efforts mandating algorithm auditing—for example, highlighting how requirements for historical preservation, central audit repositories, and standardized, accessible audit formats can strengthen audit-mandating legislation. We also contribute a dataset combining LL144 bias audit contents to date in a machine-readable format, and a codebase for extracting and evaluating future bias audit results.

3 Methods

We identified, extracted information from, and analyzed dozens of bias audit reports published by employers, tool vendors, and auditors in connection with LL144. This included both quantitative and qualitative data extraction done by a team of six researchers who collectively spent hundreds of hours locating and wrangling these bias audits due to the myriad challenges we encountered in finding, processing, and analyzing them. The outputs of our work include a dataset that combines, using a common data schema, the results of all LL144 bias audits published to our knowledge from July 2023 through November 2024.[2] Our synthesis and standardization of LL144 bias audits enabled analyses not previously possible, including analyses of bias audits across tools, auditors, data types, and years. With this paper, we also share our codebase and documentation describing our data collection and cleaning choices, which can be used by other researchers to both replicate our analysis and extract information from future LL144 bias audits. Below, we outline our process of searching for, archiving, and extracting data from the bias audits, as summarized in Figure 1.

3.1 Searching for Bias Audits

The first challenge we encountered was locating the bias audits. LL144 requires that employers make their bias audit summaries and other information "publicly available on the employment section of their website in a clear and conspicuous manner" [94]. In the audits we reviewed, employers interpreted this requirement in varying ways, and at times disregarded it. To our knowledge as of the time of writing, NYC DCWP does not maintain a public-facing repository of LL144 bias audits, so locating bias audits was very time-consuming and challenging. We conducted a thorough search process totaling hundreds of hours over more than a year; we began searching for bias audits as soon as the law took effect in July 2023 and continued to search periodically over the course of the next 18 months. In late 2023, a subset of our team created a public-facing tracker of all bias audits we had encountered, updating it as we learned of other audits [116]. We identified audits through a variety of means, including: through searches of employers' websites (especially their "Careers" pages and individual job descriptions, privacy policies, and other areas focused on regulations and compliance), searches of auditors' websites, cross-referencing our list of audits with those of other researchers, and periodic searches with specific keywords using popular search engines. We refined these keywords over time as we saw patterns emerging in audit formats. Identifying one audit on an auditors' website sometimes enabled us to find several others produced by the same auditor, a search process we dubbed "snowball searching," akin to snowball sampling in survey research or snowball searching in literature reviews [32, 47].

[2]We were unable to verify data for one audit report we collected during this time period—between our extraction and validation processes, it was removed from the website where it had been posted, and we had not properly preserved it, so we removed this audit report from our dataset.

3.2 Archiving Bias Audits

Throughout the process, we encountered issues preserving audits to retain copies for future analysis. As our data collection stretched into 2024—and recognizing that LL144 prohibits the use or continued use of an AEDT if "more than one year has passed since the most recent bias audit" [94]—we noticed a trend of employers or auditors posting an updated audit at the exact same URL as they had previously posted the initial bias audit, overwriting the initial audit. To preserve the audits we found over time, we created permanent links using Perma.cc [125] and included the links in our public-facing tracker. LL144 also does not specify what filetype or format audits should take; the bias audits we located took a variety of formats, including report-style PDFs (e.g., [70]), presentation-style PDFs (e.g., [83]), and webpages with results embedded in the HTML or included in HTML tables (e.g., [8, 63]).

3.3 Extracting Data from Bias Audits

Alongside developing a data extraction and data validation pipeline, we defined a common data schema to store audit results in a unified format as tabular data for consistent comparison and analysis. We created columns in our data schema for different types of information required to be reported in bias audit summaries under LL144 and added other columns based on patterns we noticed as we reviewed the audits. As we developed the schema, we resolved disagreements and modifications through consensus among the research team. Under our data schema, each row represents a result reported in a LL144 bias audit for a particular group or intersection of groups, and also includes columns noting the deploying entity, auditor, year, and other metadata. Some of the information included in our schema—such as the race, sex, or intersectional group for each result, as well as the number of applications, number of applicants selected or above the median score, selection rate, and impact ratio for that group—were often reported in tables in the underlying bias audits. Other information, such as the type of data used for the audit or the time frame in which the audit's data was collected, was usually noted in audits' text, footnotes, or missing altogether. As a result, our data extraction required automatic tabular data extraction and manual data entry based on reading each audit.

3.3.1 Data Extraction. For the extraction of tables, which were generally not in a machine-readable format, we implemented customized functions using the `tabulapdf` and `tabulizer` R libraries to handle formatting nuances such as multi-line headers, multi-line cells, tables stretching across multiple pages, and other audit report inconsistencies. In some instances when automated extraction proved cumbersome or ineffective, we extracted data by copying and pasting tables or manually entering the relevant information into `.csv` or `.txt` files. For all audits, we manually added information such as the deploying entity (i.e., the hiring company or vendor), position (e.g., software engineer), tool vendor (e.g., HireVue), and other details. We also kept track of how many job applications employers reported as missing demographic data, and when it was unclear whether or how much missing data there was in a given audit. Further detail on our data schema is available in our codebase documentation. Our qualitative data extraction included a "bias audit reading group," where members of the research team each read a subset of the audits and collectively discussed emerging themes between them, which informed our data extraction process and analysis. The author team also noted examples of parts of audits that made broad claims about bias, contained inconsistencies, and were particularly poorly done (informally dubbed our "auditing hall of shame") and recorded other challenges we encountered. LL144 applies to both hiring and promotion, and we designed our data schema accordingly. We did not find any audits focused only on tools used for promotion, so we focus our analyses here on hiring.

3.3.2 Data Cleaning and Validation. After extraction, we implemented processes for data cleaning and standardization, including for recording missing data, standardizing demographic data using the categories outlined in LL144 [94], and conducting data quality checks. For data validation, we manually reviewed all entries in our final dataset, cross-referencing them with the underlying audits to check the accuracy of our extraction. Throughout the extraction and validation process, we noticed what we believed to be various typos or errors by auditors (as well as other inconsistencies and ambiguities) in the bias audits we reviewed. In some instances, we felt these inconsistencies were not substantive, and adopted data entry norms that corrected these to prevent information loss. For instance, in cases where an audit report listed the month and year but not the day of the audit's data collection start date, we chose to record the day as the first of the month, noting this in our code and documentation. In other instances, the issues were more substantial, such as cases when the selection rate reported for a particular group was not equal to the number of individuals selected divided by the number of individuals who had applied, as reported in that same table. We decided to record this data as it was reported, noting these issues. We felt that trying to correct the errors or interpret the ambiguities would be out of scope for our work analyzing the audits as published, and in many instances might generate more questions than answers.

4 Findings and Discussion

Using the dataset we extracted from the bias audits—key details of which are summarized in Table 1—we conducted a mixed methods analysis of the 116 audits in our dataset. We present our results in two steps: first, we outline takeaways from each of our analyses of the bias audit results as they are reported. We analyze the type of data used, the impact ratios reported, auditors' and vendors' characterizations of audits findings, and we conduct comparisons of 2023 and 2024 bias audits of the same tool deployed by the same entity. Then, for each analysis, we highlight how further inspection of the bias audits reveals additional shortcomings, sometimes drastically altering the interpretation of bias audit findings. We demonstrate that many audits may under-report disparities in decision-making due to missing demographic data, rely on opaque data aggregation and problematic uses of "test data," and use metrics that do not faithfully represent how automated hiring tools are deployed. While a small subset of the issues we identified appear to stem from non-compliance with the rules of LL144, we identified myriad ways in which bias audits done in adherence with the law are nevertheless incomplete and potentially even inaccurate.

4.1 Data Used for Bias Audits

4.1.1 Historical Data, Test Data and Data Pooling. Under LL144, audits can use "historical data" or "test data," where "historical data" is defined as "data collected during an employer or employment agency's use of an AEDT to assess candidates for employment or employees for promotion" and "test data" is defined as "data used to conduct a bias audit that is not historical data." An auditor can use historical data pooled together from multiple employers' uses of the same AEDT (including data from different jobs or types of positions), and they can use test data "if insufficient historical data is available to conduct a statistically significant bias audit" [94]. The use of test data was relatively rare in our sample; 75% of audit reports in our sample used historical data, 16% used test data, and for 9% we were not able to identify the type of data used. However, we found serious limitations with both the use of test data and historical data. Practices related to test data seemed to vary widely, ranging from audits that generated fake applicant AEDT scores using python's `random.uniform` function (e.g., [82]) to test data that produced equal selection rates for every group in the audit (e.g., [72]). Such cases raise serious concerns about whether the bias audit results have any relationship whatsoever to the audited tool's impacts in practice. For audits that used historical data, the potential for data pooling (and the lack of transparency around it) made it extremely difficult to ascertain the impacts of particular AEDTs or employers' practices. Wide variation in how employers use automated hiring tools is common, including employers setting different thresholds that determine whether a candidate's AEDT score is high enough to advance in the hiring process [25]. In theory, under LL144, a given employer could rely on a bias audit of an AEDT they use, even though the bias audit may include very little—or no—data produced from the employer's use of the AEDT. As a reader of a bias audit posted by a given employer, it was often impossible to tell how much of that specific employer's actual data was used for the bias audit. In effect, bias audits that adhere to LL144 often fail to provide meaningful information about whether an employer's use of a given AEDT may be discriminatory.

4.2 Results Reported in Bias Audits

Next, we analyze the selection rates, impact ratios, and other details from 2023 and 2024 bias audits, highlighting how "silent duplicates" and missing data add even more challenges to attempts to identify discrimination using LL144 audits.

4.2.1 Silent Duplicates. We identified several instances where it appeared that results from the same underlying analyses (e.g., calculations of scoring or selection rates and impact ratios for a particular tool and demographic category) were reported multiple times, a phenomenon we dubbed *silent duplicates* (our definition of silent duplicates is described in more detail in Appendix A.1). We identified two types of silent duplicates—sometimes, silent duplicates appeared across different audit reports, when different employers posted audits that seemed to describe the same tools and had many identical quantitative results; we call these *inter-audit silent duplicates.* (e.g., [48] and [52]). In other instances (e.g., [49]), a given audit report included results for the same AEDT and dataset using multiple selection thresholds. Sometimes these different thresholds produced differences in selection rates for the same groups, and

Term	Our Definition	Unique Num. in our Data
Deploying entities	Organizations posting bias audits on their websites.	31
Tool vendors	Organizations selling or licensing automated tools being audited. Some vendors post audits directly, so some tool vendors are also deploying entities in our dataset.	22
Auditors	Organizations conducting bias audits of automated tools (independent from the deployer/vendor of the tool under audit).	11
Audit reports	Document(s) posted on an employer's, employment agency's, or vendor's website and characterized as a bias audit with a reference to LL144. Each audit report may contain multiple audits and is a unique combination of deploying entity, tool vendor, year, and auditor.	44
Audits	Bias audit results for a particular tool and position. More formally, each audit is a unique combination of deploying entity, vendor, year, auditor, tool ID, and employment position.	116
Multi-year audit reports	Audit reports for which results were available for both 2023 and 2024.	13

Table 1: Key details for our dataset, including definitions of terms we use throughout our analysis and the number of unique entries per term. We recorded the tool vendor or auditor as "unknown" when we could not identify this information. The most common auditors were Holistic AI (20% of audit reports), DCI Consulting Group (20%), and BABL AI (16%).

other times, they did not (we refer to these instances of identical results within the same audit as *intra-audit silent duplicates*). While we were not able to identify the cause of all the silent duplicates we observed (e.g., data pooling, shared underlying audits, coincidence), all of them appeared in audits conducted by DCI Consulting Group, and all but one described HireVue tools.

Based on our analysis, it seemed likely that bias audit reports of HireVue tools posted by JetBlue (e.g., [48]), Citizens (e.g., [52, 53]), Pfizer ([55]) and Burlington (e.g., [49, 51]) include results from the same underlying analyses. Given the data pooling allowed under LL144, it is plausible that different employers' audits of the same tools produced by the same auditor may include some identical results. But the lack of clarity on when and why silent duplicates occur further underscores how challenging it is to identify potential discrimination using these audits. While we are not aware of examples of silent duplicates beyond LL144 bias audits, we believe this issue is relevant more generally for algorithm audits in contexts where the same automated tools may be used by different entities.

4.2.2 Impact Ratio Analysis. As described in Section 2, the key statistics required for each race and/or sex group in LL144 bias audits are the *selection rate* or *scoring rate* (for selection and scoring tools, respectively) and the *impact ratio*. Analyzing the impact ratios reported, we found that 53% of audits include at least one impact ratio below 0.8 (the rate corresponding to the four-fifths rule we discussed in Section 2). When analyzing impact ratios at the group level (e.g., considering impact ratios for applicants broken down by sex and broken down by race for the same tool as separate sets of results), 23% of results across all audits included at least one impact ratio below 0.8. All of these impact ratios below 0.8 resulted from analyses broken down by race or intersectional categories (none were from analyses broken down solely by sex). Of group-level results with impact ratios below 0.8, 41% were instances where a selection or scoring rate for a group of non-White applicants that was less than four-fifths that of a group of White applicants. In the other 59% of cases, there were fewer clear trends; both the comparator group and group with a selection or scoring rate less than 0.8 varied.

There were multiple complicating factors in our impact ratio analysis that highlighted important nuances of LL144. The law permits bias audits to exclude impact ratios for groups representing less than 2% of the data being used for the bias audit [94], which frequently resulted in impact ratios not being reported for applicants who were classified as "Native Hawaiian or Other Pacific Islander" or "American Indian or Alaska Native." LL144 also defines the impact ratio as comparing a given group's selection or scoring rate against the selection or scoring rate of the highest scoring or most selected group, a standard which, if followed, would mean that impact ratios should only take on values in $[0, 1]$. However, we found that 54% of audits had at least one impact ratio *above* one, meaning that the group with the highest selection or scoring rate was *not* the comparator group in these calculations. There were various reasons why some audits had impact ratios above 1, generally resulting from differing interpretations of which groups should be included when determining the "comparator group." These instances, as well as others described in Appendix A.2, highlight how a standard that may appear to be relatively straight-forward or well-specified, such as the definition of the impact ratio, can actually be quite nuanced, complicating both interpretation and analysis of audit results. However, as we discuss in Section 4.2.4, a much more pervasive issue—missing applicant demographic data—adds even more complexity to interpreting the bias audit results.

4.2.3 Comparing Audits from 2023 to 2024. For 13 of the 25 audit reports we identified for 2023, we were also able to identify 2024 companions (listed in Appendix A.3), suggesting mixed adherence with the yearly auditing requirement. We conducted an initial comparative analysis of these audits, finding that almost all the 2024 audit reports used the same auditor and data type—historical or test data—as their 2023 predecessors, but that audit reports from 2024 widely varied in their data selection approaches compared to 2023. For instance, some audit reports (e.g., [69, 73]) appeared to use data from the same time period as the previous audit, while others used more recent data (e.g., [67, 71]) sometimes along with older data (e.g., [7, 11]). Our analysis was limited given available data; more expansive temporal analysis will be important to understanding the impacts of the law as additional LL144 bias audits are published.

4.2.4 The Impact of Missing Data. A dominant theme of our data collection and analysis was missingness—including missing information that is required to be reported under LL144, missing demographic data on large swaths of applicants, and even missing basic details about auditors or tool vendors. For example, though LL144 requires disclosure of the date of first use of an AEDT [94], we could not identify the date of first use in 86% of the audits we analyzed. Though they are not explicitly required to be reported under LL144, we also struggled to identify the job roles or positions for which the AEDT was used (this was the case for 62% of audits), and at times, the name of the auditor (for 3.4% of audits). This missingness made it challenging to analyze individual audit reports and also complicated our comparative analysis for 2023 and 2024 audit reports. For instance, without a version number or other information identifying whether a given tool had substantially changed between 2023 and 2024 audits, it was difficult to tell how changes in the selection rates and impact ratios across years should be interpreted, complicating comparisons over time.

Missing demographic data also complicated all of our analyses of bias audit results. LL144 requires bias audits to list the number of individuals who are not included in the required selection/scoring rate and impact ratio calculations because their demographic information (e.g., race and/or sex) was unknown [94]. In general, missingness was the norm—83% of audits in the sample reported that their audits were missing race and/or sex information for some applicants evaluated by the AEDT, and several bias audits reported missing demographic data for far more applicants than the number of applicants with available information.

To demonstrate the impact of missing demographic data, we calculated bounds on the impact ratios reported in bias audits when information about the number of applicants missing demographic information was available. Let x be the number of applicants missing demographic information for demographic category A (e.g., race/ethnicity, sex, or intersectional) in an audit. For each group $a \in A$, we assume that the number of applicants missing demographic information could all belong to group a, and calculate the highest and lowest possible selection rates and impact ratios for that group, holding all else equal (e.g., assuming the selection rates for other groups are the same). Let a_{app} and a_{sel} be the number of applicants with known demographic information from group a who applied to and were selected by the AEDT respectively; then the bounds of the selection rate for group a (SR_a) are thus given by Equation 1:

$$\frac{a_{sel}}{a_{app} + x} \leq SR_a \leq \frac{a_{sel} + x}{a_{app} + x} \tag{1}$$

We compute demonstrative impact ratio bounds by dividing the upper and lower selection rate bounds for group a by the selection rate of the most selected group in category A. We plot bounds for impact ratios in Figure 3 (and provide an explanation of how they can be interpreted in Figure 2), demonstrating that, while most audits that disclose the number of applications missing demographic information generally report impact ratios above 0.8, the impact ratios could often be below 0.8 depending on the distribution of the missing demographic data (the percentage of impact ratios with possible lower bounds below 0.8 is 70%). Because there were often very few or no applicants labeled "Native Hawaiian or Other Pacific Islander" or "American Indian or Alaska Native", the existence of even a small amount of missing data means that the impact ratios for these groups could take on any value in practice.

The bounds in Figures 2 and 3 are meant to be demonstrative, but we recognize that the assumption that all missing data comes from one group is unlikely to hold in practice. We explored various alternatives, including assuming that the distribution of missing data is the same as the demographic distribution of applicants with known demographics, under which the percentage of impact ratios with lower bounds below 0.8 falls to 28%. After exploring these options, we chose the approach depicted in Figures 2 and 3, in part because each option indicated that impact ratios may vary significantly from the reported values, depending on the missing demographic data. We provide more detail in Appendix A.4.

4.3 Interpreting Bias Audits

4.3.1 Bias audit characterization. In several instances, we noticed auditors, tool vendors, and employers making claims—sometimes very broad ones—about their bias audit results, even though these audits suffered from the myriad issues we identified in our analysis. For instance, one bias audit that relied purely on test data included a dashboard that used a green/yellow/red coloring system to characterize audit results, and all results for the automated tool being audited were green (corresponding to "clear") [12]. Another vendor wrote that their bias audit had "ensur[ed] that our candidate screening features are not influenced by any biases" [60], but the public facing version of this bias audit did not (to our knowledge) report the number of applicants in each group or whether any candidates evaluated by the tool were missing demographic data, and thus not included in the audit calculations. One auditor included determinations of whether a tool "passed" various audit standards—including in audits that were missing demographic data for the vast majority of applicants evaluated by the tool being audited [66, 70]. We did not identify any instances of a published audit done by this auditor that was characterized as having "failed" in any respect. As our analysis has highlighted, there are many ways these bias audits can miss signs of bias, so sweeping conclusions of any nature about these bias audits can be misleading.

5 Ten Recommendations for Future Legislation Requiring Algorithm Audits

Our analysis revealed a need for greater accessibility, transparency, and rigor in algorithm auditing than mandated by LL144. The shortcomings we found speak to problems in nearly every stage of the audit and reporting process, and as of the time of writing, we are

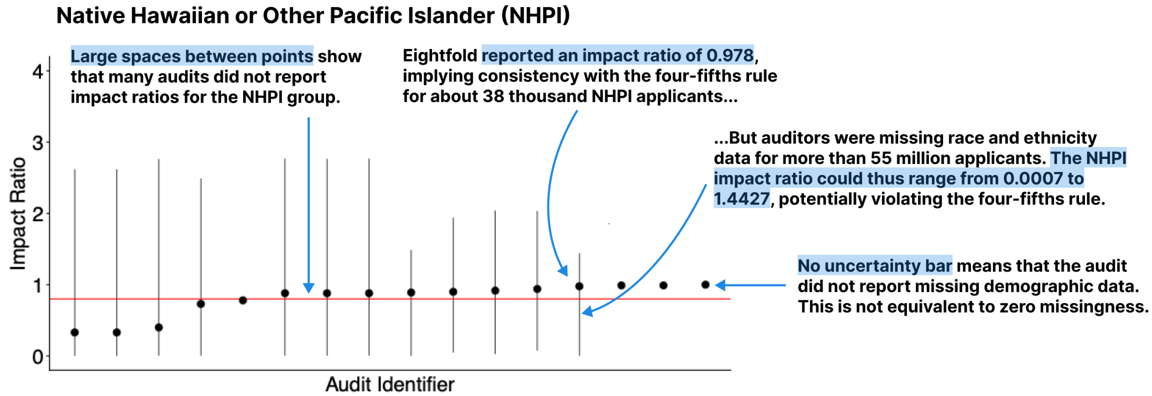

Figure 2: Impact ratios (dots) with uncertainty bounds (vertical lines) for the Native Hawaiian or Other Pacific Islander group, with annotations and examples from [70] explaining how to interpret the plot. The horizontal red line marks the four-fifths rule commonly used in disparate impact evaluations. Results for additional groups are included in Figure 3.

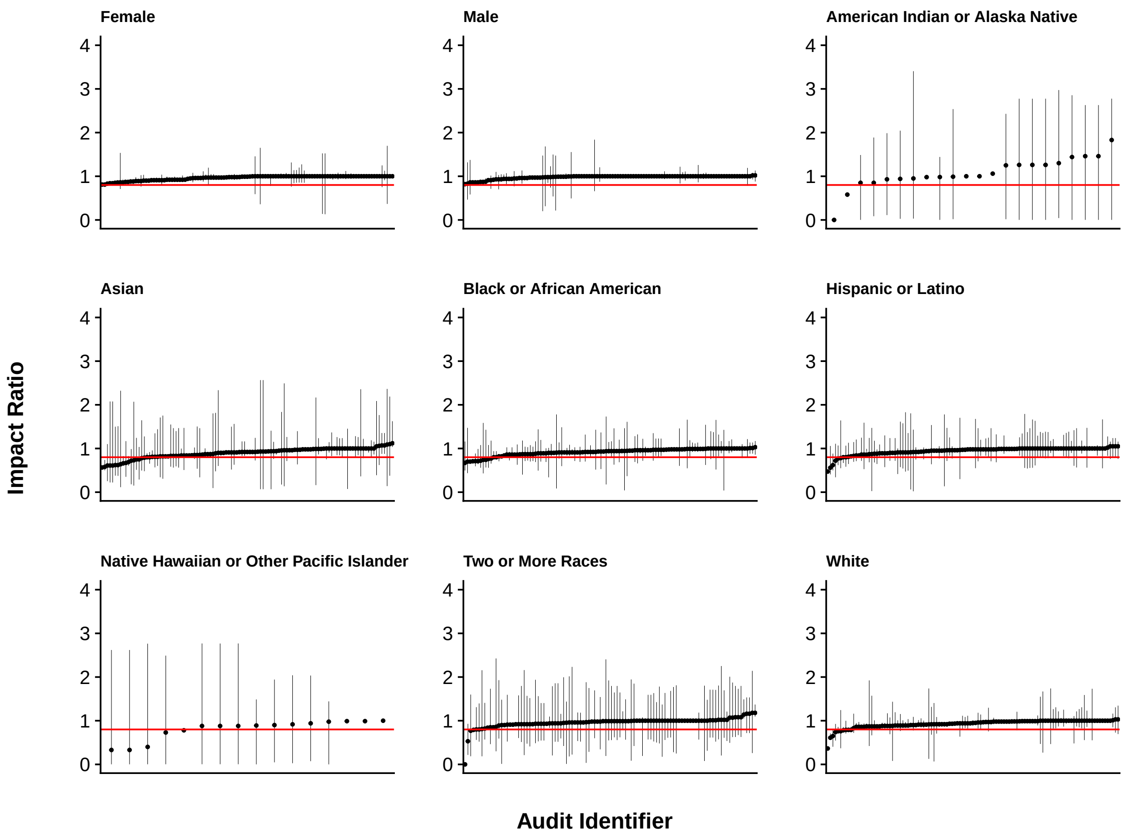

Figure 3: Impact ratios by protected characteristic. Dots represent impact ratio results from a given audit for a particular group, and vertical lines are uncertainty bounds for the impact ratios calculated from the missing demographic data. The horizontal red lines mark the four-fifths rule commonly used in disparate impact evaluations. While most audits generally report impact ratios above 0.8 (as shown by the black dots above the red line), the impact ratios reported could often be below 0.8 (as shown by the vertical lines often crossing the red line), depending on the distribution of the missing demographic data.

unaware of any enforcement actions taken by the NYC DCWP related to LL144. We propose ten recommendations, summarized in Figure 1, to strengthen future legislation that mandates algorithm auditing in hiring (though many apply beyond hiring as well) and the ensuing audits.

5.1 Recommendations Stemming from Our Search, Archival, and Extraction Processes

5.1.1 Audit Search: Central Repository of Audits and Standard URL Format. The irregular reporting and structure of LL144 bias audits made it laborious to find less than two years' worth of audits. To ensure the accessibility of audit data, audit-mandating laws should require employers to submit audits to the relevant enforcement body, and this agency should maintain a webpage where all audits are publicly accessible, providing a central location for audit data and, in effect, preventing employers from silently changing audit results. Employers should also be required to post audits in a standardized, easily accessible location on their websites, for instance, using a standard URL, e.g.: `[employer].com/[law_name]/audit/[year]`.

5.1.2 Audit Archiving: Requirements for Historical Preservation. Employers should be required to keep previous audits available and not overwrite them with new versions. Each new audit should be submitted to the relevant enforcement agency and posted at a new URL (e.g., using the standard URL format outlined above), preserving a continuous and accessible historical record of audit data.

5.1.3 Audit Extraction: Standardized Audit Format. Audits should be posted in machine-readable formats to improve the efficiency of data extraction and should align with accessibility standards (e.g., the Web Content Accessibility Guidelines [119] or similar standards) to ensure they are accessible, including for users of assistive technologies. The audits should include an identifier for the AEDT with a version number, the date of the tool's first use, and the time frame during which data used for the audit was collected or created, recorded in a structured format. For LL144 bias audits we analyzed, this information was often embedded in the text of audits or missing altogether.

5.2 Recommendations Based on Audit Analysis

5.2.1 Assessing bias for broader set of protected groups. Nearly every audit we reviewed published the selection rates and/or impact ratios for strictly the set of protected categories required by LL144. Legislation requiring algorithm audits should require auditors to assess the risk of potential bias against a broader set of groups and intersectional categories, recognizing that discrimination in hiring has long been a concern for characteristics beyond race and sex, such as disability, pregnancy, and age, among others. Legislators could look to protected categories outlined under existing civil rights laws to craft these standards; while many anti-discrimination laws do not provide explicit intersectional protections, there have been some recent court rulings and newer laws with explicitly intersectional protections (e.g., [74]). The relevant enforcement agency should establish, and auditors should use, carefully developed, standardized demographic categories and definitions to make results more comparable over time and across tools. While LL144 requires the use of the race and ethnicity categories defined by the Equal Employment Opportunity Commission's EEO Component 1 report [94], we observed variation in adherence to or interpretation of this requirement.

5.2.2 Greater auditor access. A fundamental limitation for algorithm audits is restricted access to the system they are evaluating for bias [109]. The audits we reviewed did not typically report their level of access, and LL144 does not specify an access requirement. However, as Groves et al. [59]'s interviews with LL144 auditors demonstrate, auditors often faced serious challenges accessing data they felt was needed for audits. In algorithm auditing more generally, while auditors are sometimes granted only black-box access to a system (i.e., they can only see system inputs and outputs), greater access to the underlying system can allow for more holistic auditing than black-box access alone, including by allowing auditors to inspect the system's architecture or test separate parts of the algorithm. Researchers have outlined many alternative access schemes [29]; for instance, future auditing laws could require what Casper et al. [28] call *de facto* white-box access, where deploying entities can grant auditors access through an off-site API that allows auditors to run arbitrary tests of the system without seeing proprietary model weights and parameters. Future legislative efforts mandating algorithm auditing should contemplate adopting more expansive access schemes where feasible.

5.2.3 Expanded analysis of missing demographic data. As Bogen et al. [26] wrote, "organizations cannot address demographic disparities that they cannot see." When demographic information is missing for large swaths of the applicant pool subjected to an AEDT's use—as our analysis in Section 4.2.4 revealed is often the case for LL144 bias audits—it is difficult to interpret the results. As such, missing data are a major obstacle that auditors must discuss. In addition to reporting summary statistics about applicants' missing demographic data (as required currently under LL144), audits should calculate selection or scoring rates, impact ratios, and other quantitative bias metrics among the applicants who did not disclose their demographic information (not currently required under LL144). Audits should calculate uncertainty bounds for the metrics they report that take into account missing data, as we did in Section 4.2.4. Auditors may be able to use the bias metrics for applicants with unknown demographics to tighten uncertainty bounds.

5.2.4 Transparency around data aggregation and using more representative bias metrics. As specified by LL144, auditors evaluated selection rates for selection tools and scoring rates for scoring tools, defining scoring rates as the proportion of a group scoring above the sample median. But this choice of the sample median is unrealistic; in practice, employers set and use different thresholds to determine whether an applicant moves forward in their process. As we highlight in Section 4.1.1, the data aggregation allowed under LL144 makes it challenging to assess whether an employer's AEDT may be discriminatory, in part because the aggregation ignores this threshold-setting. Researchers have highlighted the importance of evaluating automated systems using metrics that represent how thresholds are used in practice [79]. Auditors should include additional metrics that measure disparities at all points in the distribution of applicant scores, carefully designed to faithfully represent how AEDTs are actually used. Auditors should be required to provide detailed explanations about how data used for an audit was collected from deploying entities and aggregated or filtered for analysis (which could also help address confusion introduced by silent duplicates), and the legislation mandating the audits should provide clear guidance to ensure the data used is robust and representative. For instance, auditors could be required to include datasheets [41] or model cards [90] about audit data and employer-set thresholds.

5.2.5 Auditor determination guidelines. The key takeaways of algorithm audits should be legible to a variety of audiences, including policymakers, job seekers, and the general public; one way to enable this accessibility would be for auditors to include a determination of an audit. These determinations should be governed by clear guidelines that are published alongside the audits. This practice could draw on norms in other contexts; for instance, Lam et al. [81] highlight financial auditors' requirement to reach a determination for each audit. For LL144, as our analysis in Section 4.3.1 highlights, some auditors and vendors impose their own, potentially misleading, conclusions in the absence of clear guidelines for how qualitative and quantitative audit results should be converted to final determinations.

5.3 Structural Recommendations to Improve Audits

5.3.1 Transparency on the deploying entity's governance and accountability. Evaluating internal governance is a crucial part of algorithm audits [81, 102]. At a minimum, published audit reports should include a section discussing the deploying entity's (and where relevant, the tool vendor's) governance and accountability structure, outlining who reviewed the results of the bias audit, how the deploying entity makes decisions relating to algorithmic bias—including updates based on bias audit findings—and the names and titles of the deploying entity's relevant points of contact.

5.3.2 Actionable requirements and robust enforcement. At the time of writing, we are unaware of any enforcement actions taken by the NYC DCWP related to LL144. Employers may have little incentive to conduct audits given this lack of enforcement (and because employers have wide latitude to determine whether their tools are subject to the law). Further, while NYC DCWP is charged with enforcing audit and reporting requirements, LL144 does not mandate any particular action based on the results of a bias audit. Enforcement of NYC anti-discrimination law falls to the city's Commission on Human Rights [97]. For auditing laws to be effective, the relevant enforcing agency must monitor deploying entities' compliance with the law. For instance, they should provide clear penalty structures for missing audit reports and visibly enforce those policies. Employers or other entities subject to the auditing law should be required to register a compliant audit with the enforcing agency prior to deploying a covered automated tool. Similar registries exist in other contexts, such as the Federal Audit Clearinghouse [3] and the federal AI use case inventory [95].

6 Limitations and Future Work

There are many ways to extend and strengthen this work moving forward. Our analysis covered less than two years' worth of bias audits; future work should revisit the patterns we identified with a larger sample of bias audits as they continue to be published, and we include code and documentation with this publication to aid in those efforts. In addition, due to the challenges we encountered in identifying bias audits, there are likely published bias audits that we did not find; we encourage other teams to help identify such cases and update these analyses. We also find it important to note that many of our recommendations would be challenging to implement, in part because they may require consensus among policymakers about open challenges in the field of algorithm auditing. But as our analysis demonstrates, meaningfully assessing algorithmic bias requires grappling with these challenges.

Finally, even if implemented fully, our recommendations are likely insufficient to ensure algorithm audits are effective. As new legislation is written, so too will new challenges emerge. While out of the scope of this work, there are many other important considerations for strengthening algorithm audits, such as ensuring the definition of covered algorithms is appropriately expansive, assessing bias based on proxies for protected characteristics, and conducting audits that examine AEDTs as they are embedded in the full hiring process—including examining interactions with other AEDTs and human decision-makers—rather than examining specific algorithm-aided decision points in a silo. There are also important considerations related to the broader audit ecosystem that are out of scope of this work, including ensuring auditor independence through licensing or accreditation systems, developing structures to prevent deploying entities from "opinion shopping" among different auditors, auditing requirements that leverage counterfactual offline and online auditing methods (e.g., [23, 44, 106, 118]) and creating incentives for auditors to uncover and disclose bias issues (e.g., through algorithmic bias bounties, akin to the idea of bug bounties in computer security [45]).

7 Conclusion

In theory, algorithm audits that are designed to mitigate or prevent algorithmic bias should not *themselves* need to be audited. But in practice, we encountered a strong need for this accountability practice while undertaking an extensive analysis of bias audits produced pursuant to one of the first enacted hiring-related algorithmic decision-making laws in the country, New York City's Local Law 144. We encountered numerous issues at each stage of our process of searching for, archiving, extracting information from, and analyzing bias audits produced pursuant to Local Law 144, painting a picture of these audits as incomplete, misleading, and potentially even inaccurate. Our analysis demonstrated that these issues affect the integrity, interpretability, and ultimate usefulness of these algorithm audits, falling short of the legislation's goals. In light of our findings, we offer ten recommendations to strengthen future auditing efforts of automated tools used in hiring and other areas, and contemplate broader issues related to the auditing ecosystem that should be further explored in future work. We hope that this work will help strengthen future auditing legislation towards audits that meaningfully prevent, identify and address algorithmic bias.

Acknowledgments

We would like to thank several of our colleagues at the ACLU and the NYCLU for helpful feedback and advice at various stages of this project, including Daniel Schwarz, Amreeta Mathai, Beth Haroules, Malika Mohan, Daniel Kahn-Gillmor, Olga Akselrod, ReNika Moore, Cody Venzke, and Brooke Madubuonwu.

References

[1] ADP. 2023. Candidate Relevancy FAQ's. https://perma.cc/KF5Y-5XY7
[2] ADP. 2024. Candidate Relevancy FAQ's. https://perma.cc/XFY4-BPSC
[3] U.S. Government Services Agency. 2025. The Federal Audit Clearinghouse. https://www.fac.gov/
[4] Conductor AI. 2023. AEDT Bias Audit - Covey. https://perma.cc/WMG2-N8HH
[5] Holistic AI. 2023. Audit Results. https://perma.cc/8MA2-P6VB
[6] Holistic AI. 2023. Bryq – Bryq Talent Intelligence Summary of Results. https://perma.cc/NNR6-9DW3
[7] Holistic AI. 2023. Paradox Summary of Results. https://perma.cc/SK87-ZFY5
[8] Holistic AI. 2023. Summary of Results. https://perma.cc/CSA8-S2RC
[9] Holistic AI. 2024. Bryq – Bryq Talent Intelligence Summary of Results. https://perma.cc/3TJD-EQE4
[10] Holistic AI. 2024. Model Audit Results. https://perma.cc/GZY5-UKZN
[11] Holistic AI. 2024. Paradox Summary of Results. https://perma.cc/CE76-L7L4
[12] Warden AI. 2024. Beamery AI Assurance Dashboard. https://perma.cc/EWS3-JUBD
[13] Ifeoma Ajunwa. 2020. An Auditing Imperative for Automated Hiring Systems. *Harvard Journal of Law & Technology (Harvard JOLT)* 34, 2 (2020), 621–700. https://jolt.law.harvard.edu/assets/articlePDFs/v34/5.-Ajunwa-An-Auditing-Imperative-for-Automated-Hiring-Systems.pdf
[14] Ifeoma Ajunwa. 2023. *The Quantified Worker: Law and Technology in the Modern Workplace.* Cambridge University Press, Cambridge. doi:10.1017/9781316888681
[15] Olga Akselrod and Ricardo Mimbela. 2024. The Long History of Discrimination in Job Hiring Assessments. https://www.aclu.org/news/racial-justice/the-long-history-of-discrimination-in-job-hiring-assessments
[16] Olga Akselrod and Cody Venzke. 2023. How Artificial Intelligence Might Prevent You From Getting Hired. https://www.aclu.org/news/racial-justice/how-artificial-intelligence-might-prevent-you-from-getting-hired
[17] Elham Albaroudi, Taha Mansouri, and Ali Alameer. 2024. A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. 5, 1 (2024), 383–404. Issue 1. doi:10.3390/ai5010019
[18] Leo Alexander III, Q. Chelsea Song, Louis Hickman, and Hyun Joo Shin. 2025. Sourcing algorithms: Rethinking fairness in hiring in the era of algorithmic recruitment. *International Journal of Selection and Assessment* 33, 1 (2025), e12499. doi:10.1111/ijsa.12499 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijsa.12499
[19] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 199 (Nov. 2019), 30 pages. doi:10.1145/3359301
[20] Idiro Analytics. 2024. Report prepared by Idiro Analytics on 7/29/2024 in accordance with NYC Local Law 144. https://perma.cc/79YC-5PJT
[21] 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.
[22] Lena Armstrong, Jayne Everson, and Amy J. Ko. 2023. Navigating a Black Box: Students' Experiences and Perceptions of Automated Hiring. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1* (Chicago, IL, USA) *(Icer '23).* Association for Computing Machinery, New York, NY, USA, 148–158. doi:10.1145/3568813.3600123
[23] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (September 2004), 991–1013. doi:10.1257/0002828042002561
[24] Emily Black, Talia Gillis, and Zara Yasmine Hall. 2024. D-hacking. 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, 602–615. doi:10.1145/3630106.3658928
[25] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias.* Report. Upturn. https://www.upturn.org/work/help-wanted/
[26] 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
[27] Lydia XZ Brown, Ridhi Shetty, and Michelle Richardson. 2020. *Report-Algorithm-driven Hiring Tools: Innovative Recruitment or Expedited Disability Discrimination.* Technical Report. Center for Democracy & Technology. https://cdt.org/insights/report-algorithm-driven-hiring-tools-innovative-recruitment-or-expedited-disability-discrimination/
[28] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. 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, 2254–2272. doi:10.1145/3630106.3659037
[29] Sarah H. Cen and Rohan Alur. 2024. From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (San Luis Potosi, Mexico) *(EAAMO '24).* Association for Computing Machinery, New York, NY, USA, Article 13, 14 pages. doi:10.1145/3689904.3694711
[30] Phoebe K. Chua, Hillary Abraham, and Melissa Mazmanian. 2021. Playing the Hiring Game: Class-Based Emotional Experiences and Tactics in Elite Hiring. *Proc. ACM Hum.-Comput. Interact.* 5, Cscw2, Article 392 (Oct. 2021), 27 pages. doi:10.1145/3479536
[31] Representative Yvette D. Clarke. 2023. H.R.5628 - 118th Congress (2023-2024): Algorithmic Accountability Act of 2023. https://www.congress.gov/bill/118th-congress/house-bill/5628/text
[32] James S. Coleman. 1958. Relational Analysis: The Study of Social Organizations with Survey Methods. *Human Organization* 17, 4 (1958), 28–36. http://www.jstor.org/stable/44124097
[33] Kevin Collier. 2023. In NYC, companies will have to prove their AI hiring software isn't sexist or racist. https://www.nbcnews.com/tech/tech-news/nyc-companies-will-prove-ai-hiring-software-isnt-sexist-racist-rcna92336
[34] 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* (Seoul, Republic of Korea) *(FAccT '22).* Association for Computing Machinery, New York, NY, USA, 1571–1583. doi:10.1145/3531146.3533213
[35] William A. Darity and Patrick L. Mason. 1998. Evidence on Discrimination in Employment: Codes of Color, Codes of Gender. *Journal of Economic Perspectives* 12, 2 (1998), 63–90. doi:10.1257/jep.12.2.63
[36] Jeffrey Dastin. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/
[37] Equal Employment Opportunity Commission (EEOC). 2022. The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees. https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence
[38] Equal Employment Opportunity Commission (EEOC). 2023. Navigating Employment Discrimination in AI and Automated Systems: A New Civil Rights Frontier. https://www.eeoc.gov/meetings/meeting-january-31-2023-navigating-employment-discrimination-ai-and-automated-systems-new/transcript
[39] Giulio Filippi, Sara Zannone, Airlie Hilliard, and Adriano Koshiyama. 2023. Local Law 144: A Critical Analysis of Regression Metrics. arXiv:2302.04119 [cs.LG] https://arxiv.org/abs/2302.04119
[40] Center for Democracy & Technology (CDT); American Association for People with Disabilities (AAPD); American Civil Liberties Union (ACLU); The Leadership Conference on Civil and Human Rights; National Women's Law Center; Upturn. 2022. Civil Rights Standards for 21st Century Employment Selection Procedures. https://cdt.org/wp-content/uploads/2022/12/updated-2022-12-05-Civil-Rights-Standards-for-21st-Century-Employment-Selection-Procedures.pdf
[41] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Nov. 2021), 86–92. doi:10.1145/3458723
[42] Marissa Gerchick and Olga Akselrod. 2023. Why Meaningful Algorithm Auditing is Key to Protecting Civil Rights in the Digital Age. https://www.aclu.org/news/racial-justice/why-meaningful-algorithm-auditing-is-key-to-protecting-civil-rights-in-the-digital-age
[43] Dave Gershgorn. 2018. Companies are on the hook if their hiring algorithms are biased. https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased
[44] Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. 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, 687–700. doi:10.1145/3630106.3658933
[45] Ira Globus-Harris, Michael Kearns, and Aaron Roth. 2022. An Algorithmic Framework for Bias Bounties. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1106–1124. doi:10.1145/3531146.3533172
[46] Ellen P. Goodman and Julia Trehu. 2022. AI Audit Washing and Accountability. *SSRN Electronic Journal* , (2022), . doi:10.2139/ssrn.4227350
[47] Leo A. Goodman. 1961. Snowball Sampling. *The Annals of Mathematical Statistics* 32, 1 (1961), 148–170. http://www.jstor.org/stable/2237615
[48] DCI Consulting Group. 2023. Airport Operations Bias Audit. https://perma.cc/6Z94-QW5E
[49] DCI Consulting Group. 2023. New York City Local Law 144 "Bias Audit" for Burlington Coat Factory for Allocation Analyst. https://perma.cc/VL7X-5T5L
[50] DCI Consulting Group. 2023. New York City Local Law 144 "Bias Audit" for HireVue for Adaptability for All Jobs. https://perma.cc/2XZG-87QM
[51] DCI Consulting Group. 2023. New York City Local Law 144 "Bias Audit" for HireVue for Adaptability for Intern or new college graduate jobs. https://perma.cc/9BK7-X4UB
[52] DCI Consulting Group. 2023. New York City Local Law 144 "Bias Audit" for HireVue for Personality - Agreeableness for All Jobs. https://perma.cc/6P58-HMMN
[53] DCI Consulting Group. 2023. New York City Local Law 144 "Bias Audit" for HireVue for Personality - Emotional Stability for All Jobs. https://perma.cc/89VD-JL47
[54] DCI Consulting Group. 2023. New York City Local Law 144 Impact Ratio Analysis for Wells Fargo for Teller. https://perma.cc/D7SW-K7WQ
[55] DCI Consulting Group. 2023. New York City Local Law 144 "Bias Audit" for HireVue. https://perma.cc/4EQP-MLYA
[56] DCI Consulting Group. 2024. Airport Operations Bias Audit. https://perma.cc/8T9A-Y2NG
[57] DCI Consulting Group. 2024. New York City Local Law 144 "Bias Audit" for HireVue for Adaptability for All Jobs. https://perma.cc/L823-M5RQ
[58] DCI Consulting Group. 2024. New York City Local Law 144 "Bias Audit" for HireVue for Adaptability for Intern or new college graduate jobs. https://perma.cc/RD7H-V44G
[59] 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
[60] Mithun Gunalan. 2024. Sense Passes Bias Audit to Comply with NYC Hiring Law. https://perma.cc/M8DB-N6XX
[61] Airlie Hilliard, Ayesha Gulley, Adriano Koshiyama, and Emre Kazim. 2024. Bias Audit Laws: How Effective Are They at Preventing Bias in Automated Employment Decision Tools? *International Review of Law, Computers & Technology* , (2024), 1–17. doi:10.1080/13600869.2024.2403053
[62] HiredScore. 2023. HiredScore's Focus on Compliance and Bias Mitigation. https://perma.cc/96ZU-U5DE
[63] HiredScore. 2024. HiredScore's Focus on Compliance and Bias Mitigation. https://perma.cc/7JQV-J9RS
[64] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(Fat*'19)*. Association for Computing Machinery, New York, NY, USA, 49–58. doi:10.1145/3287560.3287600
[65] REAL Women in Trucking. 2022. Women truck drivers file complaint against Meta for algorithmic discrimination against women and older workers when delivering job ads on Facebook. https://www.realwomenintrucking.org/post/women-truck-drivers-file-eeoc-complaint-against-meta
[66] BABL AI Inc. 2023. Summary of Bias Audit Results (Audit of Eightfold's Matching Model For New York City's Local Law 144). https://perma.cc/3JGK-7H76
[67] BABL AI Inc. 2023. Summary of Bias Audit Results (Audit of RippleMatch's Fit Score Algorithm For New York City's Local Law 144). https://perma.cc/BXW6-7EMA
[68] BABL AI Inc. 2023. Summary of Bias Audit Results (HackerRank's Image Analysis System). https://perma.cc/7XQ4-UXLE
[69] BABL AI Inc. 2023. Summary of Bias Audit Results of the HackerRank's Plagiarism Detection System for New York City's Local Law 144. https://perma.cc/V5D7-42ZK
[70] BABL AI Inc. 2024. Summary of Bias Audit Results (Audit of Eightfold's Matching Model For New York City's Local Law 144). https://perma.cc/M5Q5-X9DW
[71] BABL AI Inc. 2024. Summary of Bias Audit Results (Audit of RippleMatch's Fit Score Algorithm For New York City's Local Law 144). https://perma.cc/A637-PUTT
[72] BABL AI Inc. 2024. Summary of Bias Audit Results (HackerRank's Image Analysis System). https://perma.cc/B4FS-G7QC
[73] BABL AI Inc. 2024. Summary of Bias Audit Results of the HackerRank's Plagiarism Detection System for New York City's Local Law 144. https://perma.cc/NU4S-7UD2
[74] California Legislative Information. 2024. SB-1137 Discrimination claims: combination of characteristics. https://leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=202320240SB1137&showamends=true
[75] Divij Joshi. 2021. *Algorithmic Accountability for the Public Sector: Learning from the First Wave of Policy Implementation*. Technical Report. Ada Lovelace Institute, AI Now Institute and Open Government Partnership. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/
[76] Patrick Kline, Evan K Rose, and Christopher R Walters. 2022. Systemic Discrimination among Large U.S. Employers*. *The Quarterly Journal of Economics* 137, 4 (2022), 1963–2036. doi:10.1093/qje/qjac024 arXiv:https://academic.oup.com/qje/article-pdf/137/4/1963/51053979/qjac024.pdf
[77] Patrick Kline, Evan K. Rose, and Christopher R. Walters. 2024. A Discrimination Report Card. *American Economic Review* 114, 8 (August 2024), 2472–2525. doi:10.1257/aer.20230700
[78] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689. doi:10.1073/pnas.1915768117 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1915768117
[79] Kweku Kwegyir-Aggrey, Marissa Gerchick, Malika Mohan, Aaron Horowitz, and Suresh Venkatasubramanian. 2023. The Misuse of AUC: What High Impact Risk Assessment Gets Wrong. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1570–1583. doi:10.1145/3593013.3594100
[80] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development. *Business Research* 13, 3 (2020), 795–848. doi:10.1007/s40685-020-00134-w
[81] Khoa Lam, Benjamin Lange, Borhane Blili-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)*. Association for Computing Machinery, New York, NY, USA, 1078–1092. doi:10.1145/3630106.3658957
[82] Landed and Conductor AI. 2024. AEDT Bias Audit - Landed. https://perma.cc/9GZ9-UMGZ
[83] Ambrose Consulting LLC and NYU Langone Health. 2023. Bias Audit. https://perma.cc/83UA-BVJ9
[84] BLDS LLC. 2023. Bias Audit Of Morgan Stanley's Use Of Eightfold Model For Scoring Applicants. https://perma.cc/J844-MEBQ
[85] BLDS LLC. 2024. Bias Audit Of Morgan Stanley's Use Of Eightfold Model For Scoring Applicants. https://perma.cc/A8TT-MN38
[86] Nikki Main. 2023. NYC-Based Companies Will Need to Show AI Hiring Tools Are Unbiased on Sex and Race. https://gizmodo.com/nyc-based-companies-will-need-to-audit-ai-hiring-tools-1850608051
[87] Simon McCormack and Daniel Schwarz. 2023. Biased Algorithms are Deciding Who Gets Hired. We're Not Doing Enough to Stop Them. https://www.nyclu.org/commentary/biased-algorithms-are-deciding-who-gets-hired-were-not-doing-enough-stop-them
[88] Jacob Metcalf. 2024. What federal agencies can learn from New York City's AI hiring law. https://thehill.com/opinion/technology/4360523-what-federal-agencies-can-learn-from-new-york-citys-ai-hiring-law/
[89] Helena Mihaljević, Ivana Müller, Katja Dill, Aysel Yollu-Tok, and Maximilian von Grafenstein. 2023. More or less discrimination? Practical feasibility of fairness auditing of technologies for personnel selection. *AI Soc.* 39, 5 (July 2023), 2507–2523. doi:10.1007/s00146-023-01726-w
[90] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(Fat* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. doi:10.1145/3287560.3287596
[91] Scott B. Morris and Eric M. Dunleavy (Eds.). 2016. *Adverse Impact Analysis: Understanding Data, Statistics, and Risk* (1st edition ed.). Routledge, .
[92] NYC Department of Consumer and Worker Protection. 2023. Automated Employment Decision Tools (AEDT). https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page
[93] NYC Department of Consumer and Worker Protection. 2023. Automated Employment Decision Tools: Frequently Asked Questions. https://www.nyc.gov/assets/dca/downloads/pdf/about/DCWP-AEDT-FAQ.pdf
[94] NYC Department of Consumer and Worker Protection. 2023. Notice of Adoption of Final Rule for Use of Automated Employment Decisionmaking Tools. https://rules.cityofnewyork.us/wp-content/uploads/2023/04/DCWP-NOA-for-Use-of-Automated-Employment-Decisionmaking-Tools-2.pdf

[95] U.S. Office of Management and Budget. . 2024 Federal Agency Artificial Intelligence (AI) Use Case Inventory. https://github.com/ombegov/2024-Federal-AI-Use-Case-Inventory

[96] Office of Federal Contract Compliance Programs, Department of Labor. 1978. Part 60-3—Uniform Guidelines On Employee Selection Procedures (1978). https://www.ecfr.gov/current/title-41/subtitle-B/chapter-60/part-60-3

[97] NYC Commission on Human Rights. 2025. Inside the NYC Commission on Human Rights. https://www.nyc.gov/site/cchr/about/inside-cchr.page

[98] National Conference on State Legislatures. 2024. Artificial Intelligence 2024 Legislation. https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation

[99] Amanda Ottaway. 2024. 'Everyone Ignores' New York City's Workplace AI Law. https://www.law360.com/employment-authority/articles/1808951

[100] Kelsey Purcell. 2024. 2024 Applicant Tracking System (ATS) Usage Report: Key Shifts and Strategies for Job Seekers. https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/

[101] Manish Raghavan and Pauline T Kim. 2024. Limitations of the" Four-Fifths Rule" and Statistical Parity Tests for Measuring Fairness. *Geo. L. Tech. Rev.* 8 (2024), 93.

[102] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. doi:10.1145/3351095.3372873

[103] ResumeBuilder. 2024. 7 in 10 Companies Will Use AI in the Hiring Process in 2025, Despite Most Saying It's Biased. https://www.resumebuilder.com/7-in-10-companies-will-use-ai-in-the-hiring-process-in-2025-despite-most-saying-its-biased/

[104] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. 2022. An External Stability Audit Framework to Test the Validity of Personality Prediction in AI Hiring. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2153–2193. doi:10.1007/s10618-022-00861-0

[105] Rep. Manny Rutinel, Rep. Brianna Titone, and Senator Robert Rodriguez. 2024. Consumer Protections for Artificial Intelligence. https://leg.colorado.gov/bills/sb24-205

[106] Alejandro Salinas, Amit Haim, and Julian Nyarko. 2024. What's in a name? Auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875* , (2024), .

[107] Hilke Schellmann. 2024. *The Algorithm: How AI Can Hijack Your Career and Steal Your Future.* Hurst Publishers, .

[108] Matthew U Scherer, Allan G King, and Marko J Mrkonich. 2019. Applying old rules to new tools: Employment discrimination law in the age of algorithms. *SCL Rev.* 71 (2019), 449.

[109] Daniel S Schiff, Stephanie Kelley, and Javier Camacho Ibáñez. 2024. The emergence of artificial intelligence ethics auditing. *Big Data & Society* 11, 4 (2024), 20539517241299732. doi:10.1177/20539517241299732

[110] Jason D. Schloetzer and Kyoko Yoshinaga. 2024. Algorithmic Hiring Systems: Implications and Recommendations for Organisations and Policymakers. In *YSEC Yearbook of Socio-Economic Constitutions 2023: Law and the Governance of Artificial Intelligence*, Eduardo Gill-Pedro and Andreas Moberg (Eds.). Springer Nature Switzerland, , 213–246. doi:10.1007/16495_2023_61

[111] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. . A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability. 3, 2 ( ), 100425. doi:10.1016/j.patter.2021.100425 pmid:35199067

[112] Jackie Turner. 2024. NYC's Anti-Bias AI Hiring Law Has Largely Failed So Far. https://www.legalreader.com/nycs-anti-bias-ai-hiring-law-has-largely-failed-so-far/

[113] American Civil Liberties Union. 2019. Facebook EEOC Complaints. https://www.aclu.org/cases/facebook-eeoc-complaints

[114] American Civil Liberties Union. 2023. Know Your Rights: Sex Discrimination. https://www.aclu.org/know-your-rights/sex-discrimination

[115] American Civil Liberties Union. 2024. ACLU complaint to the FTC regarding Aon Consulting, Inc. https://www.aclu.org/documents/aclu-complaint-to-the-ftc-regarding-aon-consulting-inc

[116] American Civil Liberties Union and New York Civil Liberites Union. 2023. Tracking Automated Employment Decision Tool Bias Audits. https://github.com/aclu-national/tracking-ll144-bias-audits

[117] Briana Vecchione, Karen Levy, and Solon Barocas. 2021. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.* , , 1–9.

[118] Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2023. Investigating hiring bias in large language models. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models.* , , .

[119] World Wide Web Consortium (W3C). 2024. Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/

[120] Elizabeth Anne Watkins and Jiahao Chen. 2024. The Four-Fifths Rule Is Not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro Brazil, 2024-06-03). Acm, , 764–775. doi:10.1145/3630106.3658938

[121] Kevin Wei, Carson Ezell, Nick Gabrieli, and Chinmay Deshpande. 2024. How Do AI Companies "Fine-Tune" Policy? Examining Regulatory Capture in AI Governance. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 1539–1555. doi:10.1609/aies.v7i1.31745

[122] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event Canada, 2021-03-03) *(FAccT '21)*. Acm, , 666–677. doi:10.1145/3442188.3445928

[123] 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

[124] Meg Young, Michael Katell, and P.M. Krafft. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1375–1386. doi:10.1145/3531146.3533194

[125] Jonathan Zittrain, Kendra Albert, and Lawrence Lessig. 2014. Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. *Legal Information Management* 14, 2 (2014), 88–99. doi:10.1017/S1472669614000255

[126] ZoomInfo. 2023. Disclosure Regarding Use of Automated Employment Decision Tools. https://perma.cc/YHV4-XVFC

[127] Senator Andrew Zwicker. 2019. Bill S1588. https://www.njleg.state.nj.us/bill-search/2024/S1588/bill-text?f=S2000&n=1588_I1

A Appendix

A.1 Silent Duplicates Definitions and Examples

Here, we expand on the discussion of silent duplicates from Section 4.2.1. We refer to identical audit results that appear multiple times—whether within the same audit or across different audits—as *silent duplicates*. More specifically, we define silent duplicates as rows in different audits where the tool ID, sex, race, number of applications, and type of tool are all identical and non-missing/not NA, and at least one of the following is true:

- The impact ratio, audit data start date, and audit data end date are all identical and non-missing and not NA
- If the impact ratio is missing or NA, then the audit data start and end dates are not missing or NA
- If the audit data start and end dates are missing or NA, then the number of applications is not equal to zero
- If the number of applications is 0, then:
  - Neither sex nor race are missing or unknown; AND
  - The audit data start and end dates are not missing and not NA (e.g., since there seem to be numerous cases where there are 0 applications for particular groups, we don't consider these to be silent duplicates unless there is a non-null match on the audit data start and end dates).

We also distinguish between *inter-audit silent duplicates* and *intra-audit silent duplicates*, where:

- *intra-audit silent duplicates* are silent duplicates from the same audit report (typically audits where HireVue tools were used, and except for a difference in threshold definitions, the results are exactly the same)
- *inter-audit silent duplicates* are silent duplicates from different audit reports (e.g., audit reports that we believe include at least some of the same results for the same tool(s) yet were posted by different entities).

For instance, the results for the "Female Two or More Races" group in the table titled "NYC Law Impact Ratios for the Combination of Gender and Race/Ethnicity" on page 4 of Burlington's 2023 audit of HireVue tools used for allocation analyst roles [49] are considered *intra-audit silent duplicates* under our definition, as the results are the same across threshold groupings (e.g., "Tiers 1 and 2 vs. Tier 3" and "Tier 1 vs. Tiers 2 and 3"), a phenomenon not observed for other groups in the same table. The results for the "Top and Middle Tier vs. Bottom Tier" group in the table titled "NYC Law Impact Ratios for Gender" on page 3 of Pfizer's 2023 audit of HireVue tools [55] and the results in the table titled "NYC Law Impact Ratios for Gender" on page 3 of Burlington's 2023 audit of HireVue tools used for new college grad or intern roles [51] are considered *inter-audit silent duplicates*. As we discuss in Section 4.2.1, all of the silent duplicates we identified appeared in audits conducted by a single auditor—DCI Consulting Group—and all but one described HireVue tools (there was one audit report with silent duplicates published by Wells Fargo where we could not identify the tool vendor [54].)

A.2 Impact Ratio Analysis

A.2.1 Impact Ratios Above One. We expand on our discussion in Section 4.2.2 of audits with impact ratios above one, meaning that the group with the highest selection or scoring rate was *not* the comparator group in these calculations. We found that 54% of audits had at least one impact ratio above 1, and the majority of the audits where we observed this norm were conducted by DCI Consulting Group on HireVue tools deployed by either JetBlue [48], Citizens [52], Pfizer [55], or Burlington [49]. In these audits, DCI Consulting Group provides a description of the aggregation process they used that sometimes led to the comparator group not being the most selected group, producing impact ratios above one. The only other audit where we observed impact ratios above one appeared to define the comparator group as the group with the highest selection rate among those that comprised more than 2% of the data set. As a result, when a group comprising less than 2% of the data set had a higher selection rate and the impact ratio was reported for that group, the impact ratio was above one [126].

A.2.2 Other Complications in Impact Ratio Analysis. In Section 4.2.2, we outline some of the complicating factors in our impact ratio analysis. Beyond those described in Section 4.2.2, we encountered complexities related to the interpretation of tool outcomes for some audits. For example, there were two tools in our sample (HackerRank's plagiarism detection [69, 73] and image analysis tools [68, 72], audited in both 2023 and 2024) for which selection was *not* the preferred outcome, changing the interpretation of selection rates and impact ratios, in contrast to the vast majority of AEDTs for which we had bias audits in our sample. We excluded these tools from our impact ratio analysis, but note for future work the importance of carefully reading bias audits for this type of critical directional interpretation issue.

A.3 Multi-Year Audit Reports

As discussed in Section 4.2.3, a full list of 2023 audit reports we identified with 2024 companions is included in Table 2.

A.4 Impact Ratio Bounds

In our construction of impact ratio bounds in Section 4.2.4, we assume that all missing demographic data comes from one group and hold other selection rates equal when computing impact ratio bounds. In our analysis process, we explored various assumptions, including (1) assuming all demographic data comes from one group (which we present results for in the main body of the paper); (2) assuming that the demographic distribution of missing data is the same as the demographic distribution of applications with known demographics; and (3) different assumptions about what happens to other groups' selection rates when one group's rate changes. Given our goal was to create demonstrative bounds, as noted in Section 4.2.4, we went with the chosen approach because all assumptions produced similar qualitative findings indicating that impact ratios may vary significantly from the reported values depending on missing demographic data and because our chosen assumption was the least opinionated or restrictive option we considered. To try to evaluate the likelihood of different assumptions bearing out in practice, we conducted a literature review on demographic data collection in employment contexts and LL144 specifically, which suggested that a more restrictive assumption—such as that the distribution of missing data models that of known data—was unlikely to occur in practice [21, 26, 59]. We also felt our chosen method was

Count	2023 Audit Report	2024 Companion Audit Report
1	ADP tools conducted by BLDS [1]	ADP tools conducted by BLDS [2]
2	Bryq conducted by Holistic AI [6]	Bryq conducted by Holistic AI [9]
3	Burlington's use of HireVue tools conducted by DCI Consulting Group (e.g., [51])	Burlington's use of HireVue tools conducted by DCI Consulting Group (e.g., [58])
4	Citizens' use of HireVue Adaptability tool conducted by DCI Consulting Group [50]	Citizens' use of HireVue Adaptability tool conducted by DCI Consulting Group [57]
5	Covey conducted by Conductor AI [4]	Covey conducted by Idiro Analytics [20]
6	Eightfold's Matching Model by BABL AI [66]	Eightfold's Matching Model by BABL AI [70]
7	HackerRank's Plagiarism Detection System and Image Analysis System conducted by BABL AI [68, 69]	HackerRank's Plagiarism Detection System and Image Analysis System conducted by BABL AI [72, 73]
8	HiredScore tool conducted by unknown auditor [62]	HiredScore tool conducted by unknown auditor [63]
9	JetBlue's use of HireVue tools conducted by DCI Consulting Group [48]	JetBlue's use of HireVue tools conducted by DCI Consulting Group [56]
10	Morgan Stanley's use of Eightfold's Matching Model conducted by BLDS, LLC [84]	Morgan Stanley's use of Eightfold's Matching Model conducted by BLDS, LLC [85]
11	RippleMatch's FitScore Algorithm conducted by BABL AI [67]	RippleMatch's FitScore Algorithm conducted by BABL AI [71]
12	Sheppard Mullin's use of Suited Law Assessment conducted by Holistic AI [5]	Sheppard Mullin's use of Suited Law Assessment conducted by Holistic AI [10]
13	Paradox's tools conducted by Holistic AI [7]	Paradox's tools conducted by Holistic AI [11]

Table 2: Full list of 2023 audit reports we identified with 2024 companions.

the most intuitive to explain mathematically and visually (given space constraints) than a more opinionated assumption (for instance, assumption (2) mentioned above).

Below we provide full results for assumption (2), under which the percentage of impact ratio bounds below 0.8 is 28%. In the plot below, there are some instances where the computed impact ratio bounds do not include the reported impact ratio. This result is because, as noted in Section 3, there were audits where the selection rate reported for a particular group was not equal to the number of individuals selected divided by the number of individuals who had applied, as reported in that same table. Though we also recomputed impact ratios manually, we decided to use the reported impact ratios in our analysis, because, as also noted in Section 3, we felt that trying to correct the errors or interpret the ambiguities would generate more questions than answers. For instance, in some cases the discrepancies were plausibly due to a lack of precision provided in the reported selection rates (selection rates were often rounded to the hundredths place), but in other instances the reason for the discrepancy was not clear. This finding also informed our decision to present results from assumption (1) in the main body of the paper, given the results from assumption (2) are quite complex to interpret.

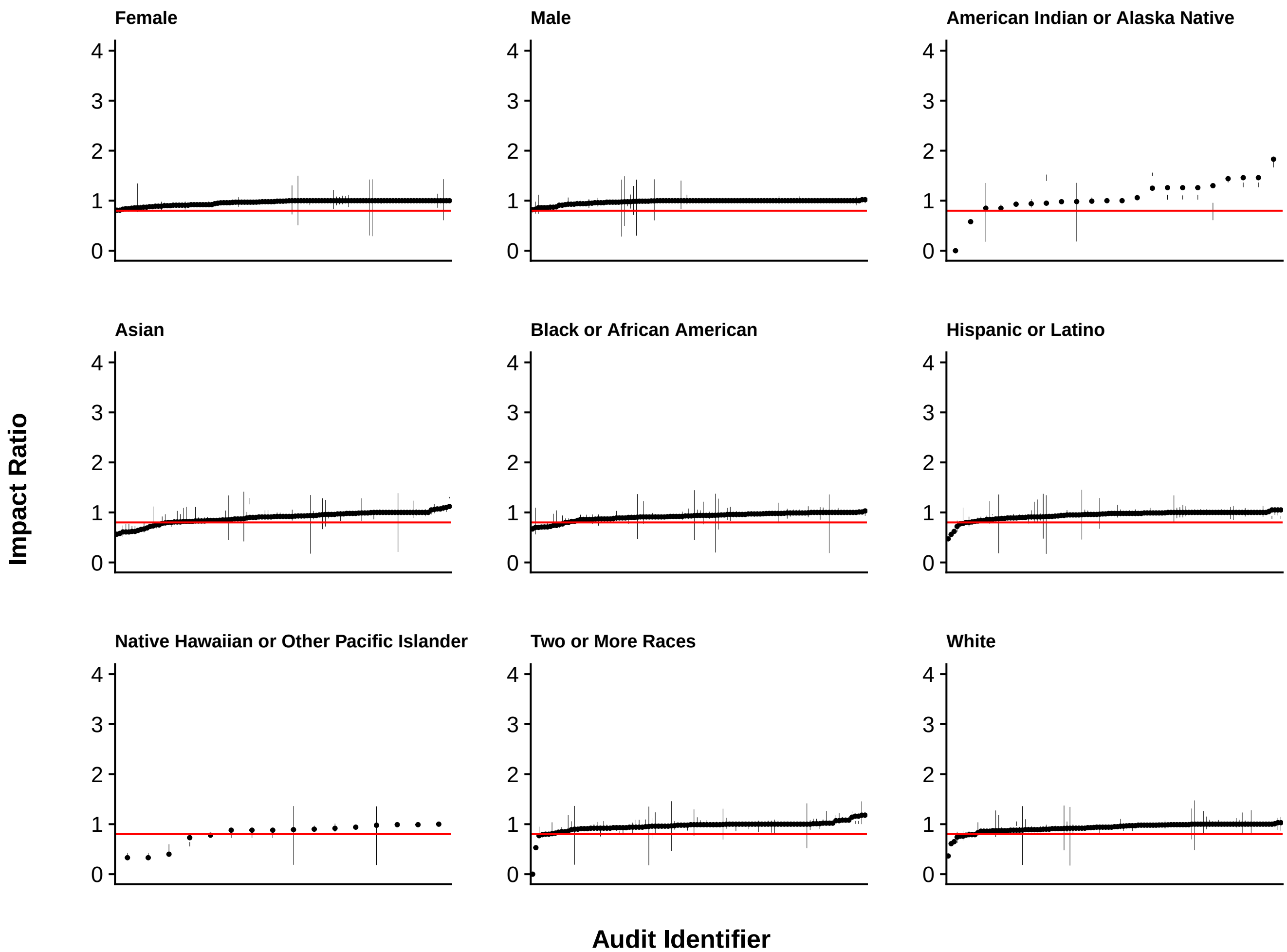

Figure 4: Impact ratios by protected characteristic under the assumption that the distribution of demographics for applications missing demographic data is the same as the demographic distribution of applications with known demographics. Dots represent impact ratio results from a given audit for a particular group, and vertical lines are uncertainty bounds for the impact ratios calculated from the missing demographic data. The horizontal red lines mark the four-fifths rule commonly used in disparate impact evaluations.