Holistic AI

Regulating the Robots:

NYC Mandates Bias Audits for AI-Driven Employment Decisions

Airlie Hilliard, Emre Kazim, Adriano Koshiyama, Markus Trengove, Sara Zannone, Nigel Kingsman, Roseline Polle.

Abstract

The mandating, by New York City, of bias audits for AI-driven employment decisions represents a significant step-change in the regulatory agenda towards greater identification and mitigation of bias in the ethically critical field of recruitment, representing a case study for broader global debates on regulating AI. In this brief, we provide an exposition of the regulation, serving two agendas: firstly providing a summary for those unfamiliar with the legislation aimed at ensuring that AI-driven recruitment tools do not prevent the hiring of a diverse workforce; and secondly, providing a commentary that discusses some shortcomings with the legislation and what it might signal for the future. Our audience is those seeking an introduction and wishing to forward the discussion via critical evaluation. **Our primary takeaways are: i. the need for examples; ii. clarification about what 'compliant' would look like; and iii. the need for a more standardised approach.**

Keywords: Bias, bias mitigation, transparency, artificial intelligence, algorithm, recruitment, psychometrics owners minimise and mitigate risk.

Introduction

The mandating, by New York City, of bias audits for AI-driven employment decisions represents a significant step-change in the regulatory agenda (Int 1894-2020, 2021). Within the broader global debates on regulating AI, it represents a real case study in a domain, namely employment, that is ethically critical. Indeed, publications have started to emerge outlining concerns about the ethical applications of algorithms in recruitment, including issues with transparency, informed consent and bias (Hunkenschroer & Luetge, 2022; Tippins et al., 2021), which echoes the wider movement toward ethical AI (Kazim & Koshiyama, 2021). In reaction to this, frameworks have been proposed to identify and mitigate bias in algorithm-driven psychological assessments (Tay et al., 2022) and to audit AI-driven recruitment systems (Kazim et al., 2021). These concerns are likely, in part, driven by increasing high-profile cases of bias in recruitment tools, such as Amazon's withdrawn AI-driven recruitment tool that penalised the resumes of female candidates containing the word "women's" due to being trained on a mostly male dataset (Dastin, 2018). In the context of displaying job adverts, Facebook and LinkedIn's algorithms were found to be biased against females, being less likely to display recruitment ads to females, particularly if the company advertising the role is male-dominated (Beatrice, 2021; Imana et al., 2021; Wall & Schellmann, 2021).

In addressing these concerns, The New York City Council has passed a local law to require bias audits for all AI-driven recruitment tools used in the city. This is an important step towards ensuring that algorithms do not present a barrier to equal employment opportunity, or prevent a diverse workforce from being hired.

In this brief, we provide an exposition of the regulation, serving two agendas:

- Firstly providing an introduction and summary for those unfamiliar with the legislative intervention; and
- Secondly, providing a commentary that situates the legislation within broader policy activity (drawing, also, on the European and UK debates).

Our intended readership is those who are interested in AI assurance and auditing, particularly in the context of recruitment, bias identification and mitigation, and equal employment opportunity. Our aim is to give readers an introduction to this legislation, which is likely to impact I-O psychologists and industry practitioners working with AI-driven recruitment tools, and to forward the discussion via critical evaluation. **Our primary takeaways concern: i. the need for examples; ii. clarification about what 'compliant' would look like; and iii. the need for a more standardised approach.**

We start by providing an overview and commentary of the statute, reproducing the legislation in italics (in quotation marks) and providing summary statements in bold. We indicate the section of the statute we are referring to by providing the section headings (indicated by §). We follow the structure of the legislation, starting with some key definitions and then moving on to the requirements, enforcement, penalties and construction of the legislation. We provide some commentary in each section and highlight some potential areas for clarification before examining how this legislation is likely a signal of what is to come in the field of AI-driven recruitment, as well as what this could signal in terms of further legislation. We reproduce the statute in its entirety in Appendix A.

Overview and Commentary

In this section, we parse the text. First, we provide an exposition of the definitions of "Automated Employment Decision Tools"; 'Bias Audit", and "Employment Decisions". We then examine the requirements, enforcement and construction of the legislation, and comment on its effective date and the implications this might have for developers and users of AI-driven recruitment tools.

Definitions (§ 20-870)

Automated Employment Decision Tools

The statute starts by defining the term automated employment decision tool. The definition has two components: one which positively defines such systems, and the other which describes what is not included. We can further parse the positive definition in terms of a technical component and a context of use.

- **Positive: […]**

•Technical component: *"computational process, derived from machine learning, statistical modeling, data analytics, or artificial intelligence, that issues simplified output, including a score, classification, or recommendation […]"*

• Context of use: *"that is used to substantially assist or replace discretionary decision making for making employment decisions that impact natural persons".*

We note here that what we have described as the technical component and context of use approach is similar to the way in which the proposed EU AI Act (European Commission, 2021) defines artificial intelligence, i.e., it is effectively any technology that assists in decision-making and has a resulting impact on a human. Whereas the EU's definition is general, here the context is delimited to employment decisions, which the statute later defines.

• Negative: *"does not include a tool that does not automate, support, substantially assist or replace discretionary decision-making processes and that does not materially impact natural persons, including, but not limited to, a junk email filter, firewall, antivirus software, calculator, spreadsheet, database, data set, or other compilation of data".*

We read the second part of the definition as a further clarification of the positive expression, insofar as it disambiguates the use of AI in recruitment from the general to the specific - for instance if a recruitment company is using AI in some other capacity (say in the context of their operations), the legislation does not apply. This signals the need for compliance across any applications of AI in the context of employment, thereby not exclusively targeting recruitment firms, but any firm using AI in their employment-relevant decisions.

Bias Audit

• Independent Auditor: *"The term "bias audit" means an impartial evaluation by an independent auditor".*

Here we read three critical dimensions:

- Bias Audit: the definition does not explicitly define what a bias audit should look like or how bias should be determined. There are multiple metrics to determine bias or adverse impact, including the Equal Employment Opportunity Commission (EEOC) four-fifths rule (EEOC, 1978) or the two standard deviations rule, which is often used by courts and is endorsed by the Federal Contract Compliance Manual (Office of Federal Contract Compliance Programs, 2020). Since different metrics can produce discrepant results (Morris & Lobsenz, 2000), particularly if sample sizes are small, the lack of clarity and standardisation surrounding which metric to use and when could be problematic, particularly for candidates who are trying to understand what different metrics mean.
- Impartial Evaluation: we read this as a point regarding methodology and standards, where the methodology should be such that it is considered reasonable and where the standards to which the audit is being conducted are considered reasonable. Concerns with this relate to the lack of precedence in the space and the lack of publicly available (and indeed codified) standards. Within the context of the UK, there has been a call for certification - this can be thought of as analogous to financial audits, where auditors must obtain appropriate certification to be qualified to conduct such audits. Indeed, in the UK at least, efforts toward AI assurance frequently make comparisons with financial auditing (Centre for Data Ethics and Innovation, 2021).

• **Nature of Audit:** *"Such bias audit shall include but not be limited to the testing of an automated employment decision tool to assess the tool's disparate impact on persons of any component 1 category required to be reported by employers pursuant to subsection (c) of section 2000e-8 of title 42 of the United States code as specified in part 1602.7 of title 29 of the code of federal regulations."*

- Other than the requirement to test for disparate impact during the audit, the statute leaves it to the discretion of the firm to decide whether there are any other aspects of the tool that they wish to audit; in our prior work, we have also proposed that audits could consider transparency (in the governance and decision-making procedures, and system explainability), safety or robustness (accurate when applied in different contexts or to different datasets), and privacy (processing data in ways that reveal the nature of the data the model was trained on or that they are utilizing when they provide an output) (Kazim et al., 2021; Koshiyama et al., 2021). This could lead to differences in levels of compliance, with some firms just doing the minimum requirements to comply with the law while others go above and beyond with their audits, also examining the verticals of transparency, safety or robustness, and privacy, as we suggest in our prior work (Kazim et al., 2021). The level of compliance is, therefore, likely to vary between businesses and might be influenced by factors such as the size of the business and the availability of resources, including financial.
- In guiding investigations of bias, the statute refers to protected attributes in the component 1 category, which refers to characteristics such as age, sex, and race, as well as familial and marital status. Since some of these characteristics are not typically collected from candidates, it is not clear from the legislation whether using the available information in bias audits is sufficient or whether firms should collect this additional information to perform a more comprehensive audit. If the latter approach is required, this could increase the risk surrounding privacy, and so compliance with privacy and data regulation legislation would also have to be considered at a deeper level. Further, it is often not possible to eliminate bias for all subgroups since mitigating bias for one subgroup can inflate it for another (Ryan et al., 1998).
- In terms of reporting the results of the audit, the legislation refers to the investigations sub-section of the Public Health and Welfare chapter in the United States Code *("subsection (c) of section 2000e-8 of title 42 of the United States code")*, which gives guidance on the execution, retention and preservation of records (U.S. Code, 2017). It also refers to the Requirement for filing of report guidance in the Record Keeping and Reporting Requirements subsection of the code of federal regulations *("part 1602.7 of title 29 of the code of federal regulations")*, which apply to firms with over 100 employees (Federal Register, 1991).

We have already seen some action towards auditing in the recruitment context, with pymetrics and HireVue publishing results of audits of their systems. The audit of pymetrics carried out by a team of computer science researchers at Northeastern University examined their compliance with the four-fifths rule and tested the robustness of their adverse impact testing framework through a code audit (Wilson et al., 2021). In contrast, HireVue hired a commercial consultancy (O'Neil Risk Consulting and Algorithmic Auditing) to carry out an audit of a pre-built assessment, which took the form of discussions with relevant stakeholders to investigate anti-discrimination practices. Based on the legislation, it is not clear what form the audit should take and whether either of these approaches is acceptable, or whether one would be favoured. To facilitate a more standardised approach to these audits, it would be beneficial to provide examples of what a desirable audit looks like.

We have previously offered guidance on algorithm auditing and how these audits can be used to mitigate potential weaknesses in the system (Koshiyama et al., 2021) and proposed a framework for auditing recruitment systems specifically. Here we take a holistic approach, considering levels of access (depth of audit), the point at which an audit should take place (locating when an audit can take place and how that point in a process impacts the nature of the audit that can be performed) and critically cover technical verticals, including bias, as well as a route to what would be appropriate documentation (Kazim et al., 2021). We summarise this framework in Appendix B, offering a resolution to the lack of definition of a standardised approach in the legislation and additional verticals that firms may wish to audit to go above and beyond the legislation. Indeed, it is not clear why the legislation is only concerned with auditing for bias; recruitment systems, which are inherently high-risk, can pose risks relating to multiple verticals.

Employment Decision

• **Employment decision.** *"The term "employment decision" means to screen candidates for employment or employees for promotion within the city".*

The statute's jurisdiction is limited to New York City, with any AI-driven recruitment tool within the city limits being subject to compliance. However, the legislation does not make it clear whether this also applies to New York City based firms who are hiring candidates outside of the city limits, or whether this applies to remote workers, whose headquarters are registered as being in New York City. Such ambiguity may result in issues with compliance unless this is clarified.

Requirements for Automated Employment Decision Tools (§ 20-871)

In this section, we expand upon the statements of the legislation that note requirements to satisfy the statute. We do this principally by sequencing the text and explaining our reading where necessary. We provide commentary on each of these points, suggesting some areas for further clarification.

• ***Evidence. "a. In the city, it shall be unlawful for an employer or an employment agency to use an automated employment decision tool to screen a candidate or employee for an employment decision unless:"***

- ***Period/Time:*** "1. Such tool has been the subject of a bias audit conducted no more than one year prior to the use of such tool; and"

- ***Documentation:*** "2. A summary of the results of the most recent bias audit of such tool as well as the distribution date of the tool to which such audit applies has been made publicly available on the website of the employer or employment agency prior to the use of such tool."

In terms of providing evidence of the bias audit, the legislation creates some ambiguity around how often these tools need to be audited. While it does specify that an audit needs to take place no more than a year before the tool is used, it is not clear whether this refers to the first time the tool is used (or the first time it is used after the legislation comes into effect), or whether this refers to its continued use. The former would mean that only one-off audits are required while the latter would require periodic audits. We take the view that because the algorithms used in these tools are often continuously updated (Tippins et al., 2021), such systems appeal to the need for continuous monitoring due to both the risk of bias being introduced or exacerbated when these updates are made, and because these alterations could impact the robustness of the model.

Further, the required format of the documentation is unclear. Considering bias metrics can be technical, and therefore lack interpretability to anyone unfamiliar with them, the target audience of these reports should be made clearer. If they are to help candidates better understand the tool and how someone in their demographic group may fare with them, which we suspect that they are, then greater guidance is needed on how these metrics can be translated into a user-friendly format to help potential candidates make an informed decision without burdening them with having to digest highly technical information. Providing examples of desirable reports is likely to benefit firms and ensure a more standardised approach to reporting.

• **Transparency Obligations.** *"b. Notices required. In the city, any employer or employment agency that uses an automated employment decision tool to screen an employee or a candidate who has applied for a position for an employment decision shall notify each such employee or candidate who resides in the city of the following:*
- 1. That an automated employment decision tool will be used in connection with the assessment or evaluation of such employee or candidate that resides in the city. Such notice shall be made no less than ten business days before such use and allow a candidate to request an alternative selection process or accommodation;
- 2. The job qualifications and characteristics that such automated employment decision tool will use in the assessment of such candidate or employee. Such notice shall be made no less than 10 business days before such use; and
- 3. If not disclosed on the employer or employment agency's website, information about the type of data collected for the automated employment decision tool, the source of such data and the employer or employment agency's data retention policy shall be available upon written request by a candidate or employee. Such information shall be provided within 30 days of the written request. Information pursuant to this section shall not be disclosed where such disclosure would violate local, state, or federal law, or interfere with a law enforcement investigation."

We can reproduce the above as follows:
• All candidates must be made aware that AI is being used at least 10 days before they are evaluated.
- While we endorse candidates being given adequate time to consider the implications AI might have for them, it is also worth considering the impact this might have on hiring timelines since the previously rapid screening of applicants might be delayed by this requirement.

• All candidates must know the characteristics which are being used in the model at least 10 days before it is used to evaluate them.
- While these obligations are a step in the right direction in providing greater transparency, there are still some shortcomings in terms of how useful this information is to candidates. Although knowing the features being fed into the model allows more informed consent, candidates still do not know how these features are being used, the outcomes they relate to, and how they interact within the model. Since these algorithms can be hard to interpret due to their often black-box nature, it is unlikely that candidates will be able to make the link between the input features and the output they result in. Therefore, although this is a step towards transparency, it is unlikely that these complex systems will ever be fully transparent.

• All candidates must be given an option to request using an alternative assessment methodology or to have accommodations.
- The legislation does not make it clear what happens if firms fail to comply with these requests.

• If candidates are not given adequate information about the features used in the model at least 10 days before the tool is used, then they must be provided with information concerning the source of the data and the data retention policy within 30 days of a written request unless providing this information would violate local, state or federal law.
- In our view, this suggests that if the firm does not disclose information about the features being used in the model prior to the candidate being assessed, then they have a second opportunity to provide this information after the fact. This might result in firms only supplying this information when requested, limiting the impact of the legislation.

Penalties (§ 20-872)

• **Sanctions.** *"a. Any person that violates any provision of this subchapter or any rule promulgated pursuant to this subchapter is liable for a civil penalty of not more than $500 for a first violation and each additional violation occurring on the same day as the first violation, and not less than $500 nor more than $1,500 for each subsequent violation*

- b. Each day on which an automated employment decision tool is used in violation of this section shall give rise to a separate violation of subdivision a of section 20-871.

- *c. Failure to provide any notice to a candidate or an employee in violation of paragraphs 1, 2 or 3 of subdivision b of section 20-871 shall constitute a separate violation.*

- d. A proceeding to recover any civil penalty authorized by this subchapter is returnable to any tribunal established within the office of administrative trials and hearings or within any agency of the city designated to conduct such proceedings.

The issuing of fines of between $500 and $1500 for each violation is not likely to be an impactful deterrent of lack of compliance. Considering that many of the firms subject to this legislation are multi-million dollar companies, a fine of this magnitude is a drop in the ocean. Perhaps more impactful might be the reputational damage that a firm might incur if they fail to comply, potentially resulting in a loss of loyalty and poor public image. However, a reason for the fines being so small might be because the sanctions relate to individual cases of violation. In other words, the fine is multiplied by the number of candidates affected by the lack of compliance, which would result in a much more costly fine. It is unclear, in our opinion, how a violation is defined or quantified.

Enforcement (§ 20-873)

• **Court proceedings.** *"The corporation counsel or such other persons designated by the corporation counsel on behalf of the department may initiate in any court of competent jurisdiction any action or proceeding that may be appropriate or necessary for correction of any violation issued pursuant this subchapter, including mandating compliance with the provisions of this chapter or such other relief as may be appropriate."*

Here the legislation indicates that the Council of New York City may initiate proceedings against any firms that fail to comply, and that they can do so through any appropriate route or in any appropriate court.

Construction (§ 20-874)

• ***Civil Proceedings Not Prevented.*** *"The provisions of this subchapter shall not be construed to limit any right of any candidate or employee for an employment decision to bring a civil action in any court of competent jurisdiction, or to limit the authority of the commission on human rights to enforce the provisions of title 8, in accordance with law."*

We read this section as an indication that the legislation does not aim to prevent or limit the ability of individuals to bring about legal action against an employer who they believe is acting unlawfully or limit the ability of the commission on human rights to enforce other relevant laws. In other words, it does not overwrite other existing laws.

Effective Date (§ 2)

• ***Effective Date.*** *"This local law takes effect on January 1, 2023."*

Considering the law was only enacted in December of 2021, we question whether firms will have enough time to ensure that they are compliant with the legislation when it comes into effect, particularly since there is a lack of guidance on what the report should look like and who is qualified to conduct the audits. We speculate that there may be a period of grace where fines are not enforced to allow firms to prepare, as was seen with the general data protection regulation in the European Union.

Discussion

We suspect that the passing of this legislation is a signalling of what is to come in the United States and indeed across the world. While the scope of this legislation is limited in terms of its enforcement being limited to the confines of New York City, we speculate that this will cause a chain reaction, where other cities and states adopt similar approaches. Indeed, although not exactly equivalent, Illinois introduced the Artificial Intelligence Video Interview Act, which requires providers to inform candidates that video interviews will be used and the features that will be analysed by artificial intelligence (820 ILCS 42, 2020). We suspect that compliance with the mandatory audits within New York City will have a ripple effect, resulting in firms using the same or similar algorithms outside of New York City also opting to audit these. It is also possible that, due to this and other forthcoming legislation, I-O psychologists will be required to carry out some audits of systems that they are not affiliated with, or that they might assist engineers in translating technical reports into more user-friendly guides for candidates. This could result in I-O psychologists needing to familiarise themselves with such regulations, burdening them with understanding the often confusing and complex wording of the legislation.

Conclusion

The mandating of bias audits for AI-driven recruitment tools is a step towards ensuring that these tools are applied in an ethical way and addressing concerns about their use (Hunkenschroer & Luetge, 2022; Tippins et al., 2021). Within the ethically critical domain of recruitment, the removal of barriers to equal employment opportunity and hiring a diverse workforce is essential. AI-driven recruitment tools provide an opportunity to do this, providing that they are trained appropriately, since it is arguably easier to mitigate bias in a well-trained algorithm than unconscious biases in humans, which are notoriously difficult to address (Equality and Human Rights Commission, 2018). Efforts have been made toward providing ways of identifying and mitigating bias in algorithmic assessments (Kazim et al., 2021; Tay et al., 2022) but the mandating of auditory bias audits represents a significant step-change in the regulatory agenda. In this article, we examined the New York City mandating of bias audits of AI-driven recruitment tools, highlighting the importance of bias audits as well as discussing potential limitations with the legislation, including a lack of clarity, a lack of standardisation and little information about what a desirable audit might look like, as well as a lack of standardisation. To address this, we have proposed a framework that can be used to carry out these audits, based on our previous work. Our key takeaways are that this mandating might see the role of I-O psychologists and indeed the regulation of AI-driven recruitment tools shift but only if the necessary examples of compliance and documentation are provided to facilitate this.

References

Beatrice, A. (2021, June 25). LinkedIn Coughed Out AI Bias! Is AI in Recruitment Reliable? IndustryWired. https://industrywired.com/linkedin-coughed-out-ai-bias-is-ai-in-recruitment-reliable/

Centre for Data Ethics and Innovation. (2021). The roadmap to an effective AI assurance ecosystem. https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women . https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290–38315.

Equality and Human Rights Commission. (2018). Unconscious bias training: An assessment of the evidence for effectiveness.

European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Federal Register. (1991). 29 CFR § 1602.7—Requirement for filing of report. LII / Legal Information Institute. https://www.law.cornell.edu/cfr/text/29/1602.7

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. Journal of Business Ethics. https://doi.org/10.1007/s10551-022-05049-6

Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for Discrimination in Algorithms Delivering Job Ads. Proceedings of the Web Conference 2021, 3767–3778. https://doi.org/10.1145/3442381.3450077

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. Patterns, 2(9), 100314. https://doi.org/10.1016/j.patter.2021.100314

Kazim, E., Koshiyama, A. S., Hilliard, A., & Polle, R. (2021). Systematizing Audit in Algorithmic Recruitment. Journal of Intelligence, 9(3), 46. https://doi.org/10.3390/jintelligence9030046

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S., & Lomas, E. (2021). Towards algorithm auditing: A survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3778998

Morris, S. B., & Lobsenz, R. B. (2000). Significance tests and confidence intervals for the adverse impact ratio. Personnel Psychology, 53(1), 89–111. https://doi.org/10.1111/j.1744-6570.2000.tb00195.x

Office of Federal Contract Compliance Programs. (2020). Federal Contract Compliance Manual (p. 527).

Ryan, A. M., Ployhart, R. E., & Friedel, L. A. (1998). Using personality testing to reduce adverse impact: A cautionary note. Journal of Applied Psychology, 83(2), 298–307. https://doi.org/10.1037/0021-9010.83.2.298

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. Advances in Methods and Practices in Psychological Science, 5(1), 251524592110613. https://doi.org/10.1177/25152459211061337

A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools, Int 1894-2020 (2021). https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced&Search

Tippins, N., Oswald, F., & McPhail, S. M. (2021). Scientific, Legal, and Ethical Concerns About AI-Based Personnel Selection Tools: A Call to Action. Personnel Assessment and Decisions, 7(2). https://doi.org/10.25035/pad.2021.02.001

References

U.S. Code. (2017). 42 U.S. Code § 2000e–8—Investigations. LII / Legal Information Institute. https://www.law.cornell.edu/uscode/text/42/2000e-8

Wall, S., & Schellmann, H. (2021). LinkedIn's job-matching AI was biased. The company's solution? More AI. MIT Technology Review. https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 666–677. https://doi.org/10.1145/3442188.3445928

Appendix A

The New York Mandatory Bias Audit Legislation

Int. No. 1894-A

By Council Members Cumbo, Ampry-Samuel, Rosenthal, Cornegy, Kallos, Adams, Louis, Chin, Cabrera, Rose, Gibson, Brannan, Rivera, Levine, Ayala, Miller, Levin and Barron

A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools

Be it enacted by the Council as follows:

Section 1. Chapter 5 of title 20 of the administrative code of the city of New York is amended by adding a new subchapter 25 to read as follows:

Subchapter 25

Automated Employment Decision Tools

§ 20-870 Definitions. For the purposes of this subchapter, the following terms have the following meanings:

Automated employment decision tool. The term "automated employment decision tool" means any computational process, derived from machine learning, statistical modeling, data analytics, or artificial intelligence, that issues simplified output, including a score, classification, or recommendation, that is used to substantially assist or replace discretionary decision making for making employment decisions that impact natural persons. The term "automated employment decision tool" does not include a tool that does not automate, support, substantially assist or replace discretionary decision-making processes and that does not materially impact natural persons, including, but not limited to, a junk email filter, firewall, antivirus software, calculator, spreadsheet, database, data set, or other compilation of data.

Bias audit. The term "bias audit" means an impartial evaluation by an independent auditor. Such bias audit shall include but not be limited to the testing of an automated employment decision tool to assess the tool's disparate impact on persons of any component 1 category required to be reported by employers pursuant to subsection (c) of section 2000e-8 of title 42 of the United States code as specified in part 1602.7 of title 29 of the code of federal regulations.

Employment decision. The term "employment decision" means to screen candidates for employment or employees for promotion within the city.

§ 20-871 Requirements for automated employment decision tools. a. In the city, it shall be unlawful for an employer or an employment agency to use an automated employment decision tool to screen a candidate or employee for an employment decision unless:

1. Such tool has been the subject of a bias audit conducted no more than one year prior to the use of such tool; and
2. A summary of the results of the most recent bias audit of such tool as well as the distribution date of the tool to which such audit applies has been made publicly available on the website of the employer or employment agency prior to the use of such tool.

b. Notices required. In the city, any employer or employment agency that uses an automated employment decision tool to screen an employee or a candidate who has applied for a position for an employment decision shall notify each such employee or candidate who resides in the city of the following:

1. That an automated employment decision tool will be used in connection with the assessment or evaluation of such employee or candidate that resides in the city. Such notice shall be made no less than ten business days before such use and allow a candidate to request an alternative selection process or accommodation;
2. The job qualifications and characteristics that such automated employment decision tool will use in the assessment of such candidate or employee. Such notice shall be made no less than 10 business days before such use; and
3. If not disclosed on the employer or employment agency's website, information about the type of data collected for the automated employment decision tool, the source of such data and the employer or

employment agency's data retention policy shall be available upon written request by a candidate or employee. Such information shall be provided within 30 days of the written request. Information pursuant to this section shall not be disclosed where such disclosure would violate local, state, or federal law, or interfere with a law enforcement investigation.

§ 20-872 Penalties. a. Any person that violates any provision of this subchapter or any rule promulgated pursuant to this subchapter is liable for a civil penalty of not more than $500 for a first violation and each additional violation occurring on the same day as the first violation, and not less than $500 nor more than $1,500 for each subsequent violation.
b. Each day on which an automated employment decision tool is used in violation of this section shall give rise to a separate violation of subdivision a of section 20-871.
c. Failure to provide any notice to a candidate or an employee in violation of paragraphs 1, 2 or 3 of subdivision b of section 20-871 shall constitute a separate violation.
d. A proceeding to recover any civil penalty authorized by this subchapter is returnable to any tribunal established within the office of administrative trials and hearings or within any agency of the city designated to conduct such proceedings.
§ 20-873 Enforcement. The corporation counsel or such other persons designated by the corporation counsel on behalf of the department may initiate in any court of competent jurisdiction any action or proceeding that may be appropriate or necessary for correction of any violation issued pursuant this subchapter, including mandating compliance with the provisions of this chapter or such other relief as may be appropriate.

§ 20-874 Construction. The provisions of this subchapter shall not be construed to limit any right of any candidate or employee for an employment decision to bring a civil action in any court of competent jurisdiction, or to limit the authority of the commission on human rights to enforce the provisions of title 8, in accordance with law.

§ 2. This local law takes effect on January 1, 2023.

SJ/ IB
LS #11903
11/1/2021 11:32 pm

Appendix B

Framework for Bias Audits of AI-driven Recruitment Tools

Based on our previous work (Kazim et al., 2021; Koshiyama et al., 2021), we present a summary of our holistic framework for the auditing of AI-driven recruitment tools. We span level of access (depth of audit), recommend appropriate points for when audits should take place (locating when an audit can take place and how that point in a process impacts the nature of the audit that can be performed) and critically cover technical verticals (bias, privacy, etc.) as well as a route to what would be appropriate documentation. We outline the risks of AI-driven recruitment tools, identify opportunities for auditing, offer some verticals that audits can relate to, including bias, examine how the level of access to algorithms can differ, and comment on assurance of these tools. For the full framework, including how contextual factors such as the level of human oversight, can influence a system, see our full framework (Kazim et al., 2021).

Risk

• Compliance risk - AI-driven recruitment tools must comply with the laws within which it is operating, including equal employment laws

•Reputational risk - companies who do not take the appropriate steps to ensure that their AI-driven recruitment tools will not result in unethical practices face the risk of reputational damage

• Financial risk - in addition to reputational damage, firms may be subject to any penalties brought about by legal action

• Governance risk - developers and deployers should be in control of and have the ability to monitor their AI-driven recruitment tools. With the risk of loss of control being increased by relying on technology, practices such as documentation and reporting can increase accountability

Auditing stages

• Data - input data should be evaluated for identifying information or protected characteristics (including proxies) while output data can be used to compute fairness and performance metrics for different subgroups

• Model - auditors could examine the model for overfitting, which threatens the generalisability of the model beyond the population it was trained on, and evaluate the model in terms of explainability and robustness

• Development - the design, building process, training routine of the algorithm and associated documentation are all processes that can be audited to examine whether there is any adverse impact analysis and the stage that this analysis takes place in Verticals

• Bias - of key interest in the context of the NYC mandatory audits is bias. There are a number of metrics that can be used to assess subgroup differences or adverse impact:

- Four-fifths rule - according to the EEOC uniform guidelines, the selection ratio of the group with the lowest selection rate should not be less than four-fifths of the group with the highest selection rate

- Statistical parity - used to determine whether a prediction is independent of a protected characteristic, where equal sub-group selection rates indicates that there is parity

- Cohen's d - used to indicate the standardised difference in the mean score of subgroups

- 2 standard deviations rule - if the standard deviation (SD) difference between the expected and observed selection rates for a subgroup is greater than 2 SDs, this indicates adverse impact

- Equal opportunity difference - the difference between the true positive rate (the proportion of candidates correctly allocated to the positive outcome, determined by checking against another model or manual allocations) of different subgroups

- Average odds difference - the average difference between the true and false positive rates of different subgroups

• Transparency - transparency in governance and decision making procedures relates to documentation and standardised assessments while transparency in terms of system explainability is concerned with the extent to which an algorithm can be explained, frequently discussed in terms of black- or white-box. When using AI-driven recruitment tools, good practice concerns being able to provide a reason why a candidate was not hired so that they can improve

• Robustness or safety - models should be accurate when applied to different contexts or to different datasets to avoid unfair decisions being made due to

a lack of model generalisability beyond the data it was trained on

• Privacy - audits of AI-driven recruitment tools should examine whether there is potential for data leakage and whether data minimisation principles are being respected, particularly since recruitment algorithms process data relating to personal and protected information

Levels of access

• The level of access an auditor has to a system ranges from white-box to black-box, where white-box systems disclose the full model details and audits of these systems have full access to the system, while black-box systems only enable indirect observation, meaning that the auditor cannot directly access the system. The former is more accurate and has the highest level of detail so for high-risk systems, this approach is more appropriate while the less detailed audit is sufficient for low-risk systems.

Assurance

• It is anticipated that by the end of the auditing process, a mechanism will be in place to assure the system. In this context, the bias audit would ensure compliance with the New York City legislation, as well as wider equal employment legislation.