Testimony of Julia Stoyanovich before the New York City Department of Consumer and Worker Protection regarding Local Law 144 of 2021 in Relation to Automated Employment Decision Tools

June 6, 2022

Dear Chair and members of the Department:

My name is Julia Stoyanovich. I hold a Ph.D. in Computer Science from Columbia University. I am an Associate Professor of Computer Science and Engineering at the Tandon School of Engineering, and an Associate Professor of Data Science at the Center for Data Science, and the founding Director of the Center for Responsible AI at New York University. In my research and public engagement activities, I focus on incorporating legal requirements and ethical norms, including fairness, accountability, transparency, and data protection, into data-driven algorithmic decision making.[1] I teach responsible data science courses to graduate and undergraduate students at NYU.[2] Most importantly, I am a devoted and proud New Yorker.

I actively participated in the deliberations leading up to the adoption of Local Law 144 of 2021[3] and have carried out several public engagement activities around this law when it was proposed [4]. Informed by my research and by opinions of members of the public, I have written extensively on the auditing and disclosure requirements of this Law, including an opinion article in the New York Times[5] and an article in the Wall Street Journal[6]. I have also been teaching members of the public about the impacts of AI and about its use in hiring, most recently by

---

[1] See https://dataresponsibly.github.io/ for information about this work, funded by the National Science Foundation through NSF Awards #1926250, 1934464, and 1922658.
[2] All course materials are publicly available at https://dataresponsibly.github.io/courses/
[3] Testimony of Julia Stoyanovich before New York City Council Committee on Technology regarding Int 1894-2020, November 12, 2020, available at https://dataresponsibly.github.io/documents/Stoyanovich_Int1894Testimony.pdf
[4] Public engagement showreel, Int 1894, NYU Center for Responsible AI, December 15, 2022 available at https://dataresponsibly.github.io/documents/Bill1894Showreel.pdf
[5] We need laws to take on racism and sexism in hiring technology, Alexandra Reeve Givens, Hilke Schellmann and Julia Stoyanovich, The New York Times, March 17, 2021, available at https://www.nytimes.com/2021/03/17/opinion/ai-employment-bias-nyc.html
[6] Hiring and AI: Let job candidates know why they were rejected, Julia Stoyanovich, The Wall Street Journal Reports: Leadership, September 22, 2021, available at https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313

offering a free in-person course at the Queens Public Library called "We are AI"[7]. Course materials are available online[8].

In my statement today I would like to make three recommendations regarding the enforcement of Local Law 144 of 2021:

1. **Auditing**: The scope of auditing for bias should be expanded beyond disparate impact to include other dimensions of discrimination, and also contain information about a tool's effectiveness - about whether a tool works. Audits should be based on a set of uniform publicly available criteria.
2. **Disclosure**: Information about job qualifications or characteristics for which the tool screens the job seeker should be disclosed to them in a manner that is comprehensible and actionable. Specifically, job seekers should see simple, standardized labels that show the factors that go into the AI's decision both before they apply and after a decision on their application is made.
3. **An informed public**: To be truly effective, this law requires an informed public. I recommend that New York City invests resources into informing members of the public about data, algorithms, and automated decision making, using hiring ADS as a concrete and important example.

In what follows, I will give some background on automated hiring systems, and will then expand on each of my recommendations.

Automated hiring systems

Since the 1990s, and increasingly so in the last decade, commercial tools are being used by companies large and small to hire more efficiently: source and screen candidates faster and with less paperwork, and successfully select candidates who will perform well on the job. These tools are also meant to improve efficiency for the job applicants, matching them with relevant positions, allowing them to apply with a click of a button, and facilitating the interview process.

In their 2018 report, Bogen and Rieke[9] describe the hiring process from the point of view of an employer as a series of decisions that form a funnel: "Employers start by *sourcing* candidates,

---

[7] "We are AI" series by NYU Tandon Center for Responsible AI and Queens Public Library helps citizens take control of tech, March 14 2022, available at https://engineering.nyu.edu/news/we-are-ai-series-nyu-tandon-center-responsible-ai-queens-public-library
[8] "We are AI: Taking control of technology", NYU Center for Responsible AI, available https://dataresponsibly.github.io/we-are-ai/
[9] Bogen and Rieke, "*Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*", Upturn, (2018) https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf

attracting potential candidates to apply for open positions through advertisements, job postings, and individual outreach. Next, during the *screening* stage, employers assess candidates—both before and after those candidates apply—by analyzing their experience, skills, and characteristics. Through *interviewing* applicants, employers continue their assessment in a more direct, individualized fashion. During the *selection* step, employers make final hiring and compensation determinations." Importantly, while a comprehensive survey of the space lacks, we have reason to believe that automated hiring tools are in broad use in all stages of the hiring process.

Despite their potential to improve efficiency for both employers and job applicants, hiring ADS are also raising concerns. I will recount two well-known examples here.

Sourcing: One of the earliest indications that there is cause for concern came in 2015, with the results of the AdFisher study out of Carnegie Mellon University[10] that was broadly circulated by the press[11]. Researchers ran an experiment, in which they created two sets of synthetic profiles of Web users who were the same in every respect — in terms of their demographics, stated interests, and browsing patterns — with a single exception: their stated gender, male or female. In one experiment, the AdFisher tool stimulated an interest in jobs in both groups, and showed that Google displays ads for a career coaching service for high-paying executive jobs far more frequently to the male group (1,852 times) than to the female group (318 times). This brings back memories of the time when it was legal to advertise jobs by gender in newspapers. This practice was outlawed in the US 1964, but it persists in the online ad environment.

Screening: In late 2018 it was reported that Amazon's AI resume screening tool, developed with the stated goal of increasing workforce diversity, in fact did the opposite thing: the system taught itself that male candidates were preferable to female candidates.[12] It penalized resumes that included the word "women's," as in "women's chess club captain," and downgraded graduates of two all-women's colleges. These results aligned with, and reinforced, a stark gender imbalance in the workforce at Amazon and other platforms, particularly when it comes to technical roles.

---

[10] Datta, Tschantz, Datta, "*Automated experiments on ad privacy settings",* Proceedings of Privacy Enhancing Technology (2015) https://content.sciendo.com/view/journals/popets/2015/1/article-p92.xml

[11] Gibbs, "*Women less likely to be shown ads for high-paid jobs on Google, study shows",* The Guardian (2015) https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

[12] Dastin, "*Amazon scraps secret AI recruiting tool that showed bias against women*", Reuters (2018) https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Numerous other cases of discrimination based on gender, race, and disability status during screening, interviewing, and selection stages have been documented in recent reports[13][14]. These and other examples show that, if left unchecked, automated hiring tools will replicate, amplify, and normalize results of historical discrimination.

Recommendation 1: Expanding the scope of auditing

Bias audits should take a broader view, going beyond disparate impact when considering fairness of outcomes. Others surely spoke to this point, and I will not dwell on it here. Instead, I will focus on another important dimension of due process that is closely linked to discrimination — substantiating the use of particular features in decision-making.

Regarding the use of predictive analytics to screen candidates, Jenny Yang states: "Algorithmic screens do not fit neatly within our existing laws because algorithmic models aim to identify statistical relationships among variables in the data whether or not they are understood or job related.[...] Although algorithms can uncover job-related characteristics with strong predictive power, they can also identify correlations arising from statistical noise or undetected bias in the training data. Many of these models do not attempt to establish cause-and-effect relationships, creating a risk that employers may hire based on arbitrary and potentially biased correlations."[15]

In other words, identifying what features are impacting a decision is important, but it is insufficient to alleviate due process and discrimination concerns. I recommend that an audit of an automated hiring tool should also include information about the job relevance of these features.

A subtle but important point is that even features that can legitimately be used for hiring may capture information differently for different population groups. For example, it has been documented that the mean score of the math section of the SAT (Scholastic Assessment Test) differs across racial groups, as does the shape of the score distribution.[16] These disparities are

---

[13] Emerging Technology from the arXiv, "*Racism is Poisoning Online Ad Delivery, Says Harvard Professor*", MIT Technology Review (2013)
https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/

[14] Stains, *"Are Workplace Personality Tests Fair?*", Wall Street Journal (2014)
http://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257

[15] Yang, "*Ensuring a Future that Advances Equity in Algorithmic Employment Decisions*", Urban Institute (2020)
https://www.urban.org/research/publication/ensuring-future-advances-equity-algorithmic-employment-decisions

[16] Reeves and Halikias "*Race gaps in SAT scores highlight inequality and hinder upward mobility*", Brookings (2017)
https://www.brookings.edu/research/race-gaps-in-sat-scores-highlight-inequality-and-hinder-upward-mobility

often attributed to racial and class inequalities encountered early in life, and are thought to present persistent obstacles to upward mobility and opportunity.

Some automated hiring tools used today claim to predict job performance by analyzing an interview video for body language and speech patterns. Arvind Narayanan refers to tools of this kind as "fundamentally dubious" and places them in the category of AI snake oil.[17] The premise of such tools, that (a) it is possible to predict social outcomes based on a person's appearance or demeanor and (b) it is ethically defensible to try, reeks of scientific racism and is at best an elaborate random number generator.

The AI snake oil example brings up a related point: that an audit should also evaluate the effectiveness of the tool. Does the tool work? Is it able to identify promising job candidates better than a random coin flip? What were the specific criteria for the evaluation, and what evaluation methodology was used? Was the tool's performance evaluated on a population with demographic and other characteristics that are similar to the New York City population on which it will be used? Without information about the statistical properties of the population on which the tool was trained (in the case of machine learning) and validated, we cannot know whether the tool will have similar performance when deployed.[18]

In my own work, I recently evaluated the validity of two algorithmic personality tests that are used by employers for pre-employment assessment[19]. This work was done by a large interdisciplinary team that included several data scientists, a sociologist, an industrial-organizational (I-O) psychologist, and an investigative journalist. My colleagues and I developed a methodology for an external audit of stability of algorithmic personality tests, and used it to audit two systems, Humantic AI and Crystal. Importantly, rather than challenging or affirming the assumptions made in psychometric testing — that personality traits are meaningful and measurable constructs, and that they are indicative of future success on the job— we framed our methodology around testing the underlying assumptions made by the vendors of the algorithmic personality tests themselves.

In our audits of Humantic AI and Crystal, we found that both systems show substantial instability on key facets of measurement, and so <u>cannot be considered valid testing instruments.</u> For example, Crystal frequently computes different personality scores if the same resume is given in PDF vs. in raw text, while Humantic AI gives different personality scores on a LinkedIn profile vs. a resume of the same job seeker. This violated the assumption that the output of a personality test is stable across job-irrelevant input variations. Among other notable findings is

---

[17] Narayanan, "*How to recognize AI snakeoil*" (2019) https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

[18] Stoyanovich and Howe, "*Follow the data: Algorithmic transparency starts with data transparency*" (2019) https://ai.shorensteincenter.org/ideas/2018/11/26/follow-the-data-algorithmic-transparency-starts-with-data-transparency

[19] An external stability audit of framework to test the validity of personality prediction in AI hiring, Rhea et al., 2022, available at https://arxiv.org/abs/2201.09151

evidence of persistent — and often incorrect —data linkage by Humantic AI. A summary of our results are presented in **Table 1**.

Facet	Crystal	Humantic
Resume file format	✗	✓
LinkedIn URL in resume	?	✗
Source context	✗	✗
Algorithm-time / immediate	✓	✓
Algorithm-time / 31 days	✓	✗
Participant-time / LinkedIn	✗	✗
Participant-time / Twitter	N/A	✓

Table 1: *Summary of stability results for Crystal and Humantic AI, with respect to facets of measurement:* ✔ *indicates sufficient rank-order stability in all traits, while* ✗ *indicates insufficient rank-order stability or significant locational instability in at least one trait, and* ***N/A*** *indicates the facet was not tested in our audit. Results are detailed in* *https://arxiv.org/abs/2201.09151*.

In summary, I recommend that the scope of auditing for bias should be expanded beyond disparate impact to include other dimensions of discrimination, and also contain information about a tool's effectiveness. To support compliance and enable a comparison between tools during procurement, these audits should be based on a set of uniform criteria. To enable public input and deliberation, these criteria should be made publicly available.

Recommendation 2: Explaining decisions to the job applicant

Information about job qualifications or characteristics that the tool uses for screening should be provided in a manner that allows the job applicant to understand, and, if necessary, correct and contest the information. As I argued in Recommendation 1, it is also important to disclose why these specific qualifications and characteristics are considered job relevant.

I recommend that explanations for job seekers are built around the popular nutritional label metaphor, drawing an analogy to the food industry, where simple, standardized labels convey information about the ingredients and production processes.[20]

[20] Stoyanovich and Howe, "*Nutritional labels for data and models*", IEEE Data Engineering BUlletin 42(3): 13-23 (2019) http://sites.computer.org/debull/A19sept/p13.pdf

An applicant-facing nutritional label for an automated hiring system should be comprehensible: short, simple, and clear. It should be consultative, providing actionable information. Based on such information, a job applicant may, for example, take a certification exam to improve their chances of being hired for this or similar position in the future. Labels should also be comparable: allowing a job applicant to easily compare their standing across vendors and positions, and thus implying a standard.

Nutritional labels are a promising metaphor for other types of disclosure, and can be used to represent the process or the result of an automated hiring system for auditors, technologists, or employers.[21]

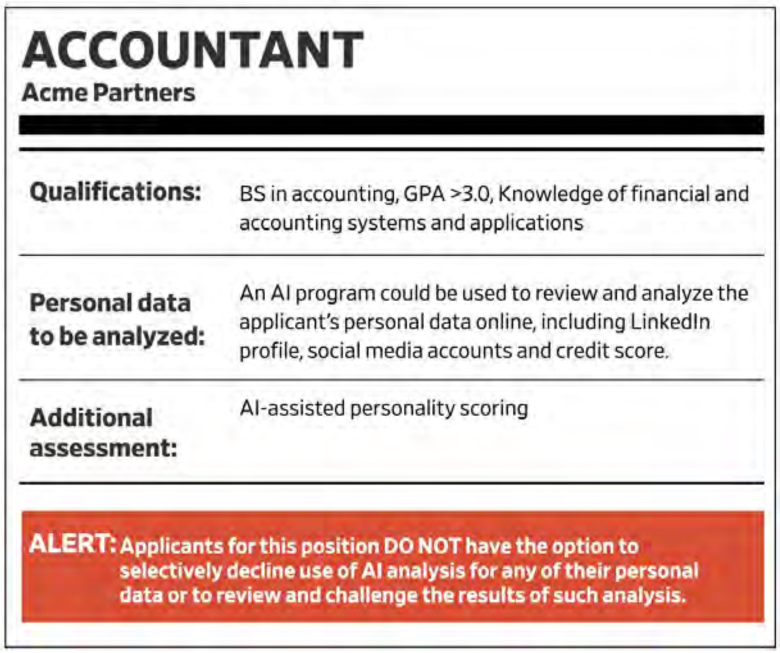

Figure 1: *A posting label is a short, simple, and clear summary of the screening process. This label is presented to a job seeker before they apply, supporting informed consent, allowing them to opt out of components of the process or to request accommodations.*

Figure 1 shows a posting label, a short and clear summary of the screening process. This label is presented to a job seeker before they apply, supporting informed consent, allowing them to opt out of components of the process or to request accommodations. Giving job seekers an opportunity to request accommodations is particularly important in light of the recent guidance

---

[21] Stoyanovich, Howe, Jagadish, *"Responsible Data Management"*, PVLDB 13(12): 3474-3489 (2020) https://dataresponsibly.github.io/documents/mirror.pdf

by the Equal Employment Opportunity Commission (EEOC) on the Americans with Disabilities Act and the use of AI to assess job applicants and employees [22].

If a job seeker applies for the job but isn't selected, then he or she would receive a"decision label" along with the decision. This label would show how the applicant's qualifications measured up to the job requirements; how the applicant compared with other job seekers; and how information about these qualifications was extracted.

Recommendation 3: Creating an informed public

My final recommendation will be brief. To be truly effective, this law requires an informed public. Individual job applicants should be able to understand and act on the information disclosed to them. In Recommendation 1, I spoke about the need to make auditing criteria for fairness and effectiveness publicly available. Empowering members of the public to weigh in on these standards will strengthen the accountability structures and help build public trust in the use of ADS in hiring and beyond. In Recommendation 2, I spoke about nutritional labels as a disclosure method. We should help job seekers, and the public at large, to understand and act upon information about data and ADS.

I recommend that New York City invests resources into informing members of the public about data, algorithms, and automated decision making, using hiring ADS as a concrete and important example. I already started this work, having developed "We are AI", a free public education course on AI and its impacts in society. This course is accompanied by a comic book series, available in English and Spanish.

Conclusion

In conclusion, I would like to quote from the recently released position statement by IEEE-USA, titled "Artificial Intelligence: Accelerating Inclusive Innovation by Building Trust".[23] IEEE is the largest professional organization of engineers in the world; I have the pleasure of serving on their AI/AS (Artificial Intelligence and Autonomous Systems) Policy Committee.

"We now stand at an important juncture that pertains less to what new levels of efficiency AI/AS can enable, and more to whether these technologies can become a force for good in ways that go beyond efficiency. We have a critical opportunity to use AI/AS to help make society more equitable, inclusive, and just; make government operations more transparent and

---

[22] The Americans with Disabilities Act and the use of software, algorithms, and AI to assess job applicants and employees, US Equal Employment Opportunity Commission, 2022, https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence

[23] IEEE-USA, "*Artificial Intelligence: Accelerating Inclusive Innovation by Building Trust*" (2020) https://ieeeusa.org/wp-content/uploads/2020/10/AITrust0720.pdf

accountable; and encourage public participation and increase the public's trust in government. When used according to these objectives, AI/AS can help reaffirm our democratic values.

If, instead, we miss the opportunity to use these technologies to further human values and ensure trustworthiness, and uphold the status quo, we risk reinforcing disparities in access to goods and services, discouraging public participation in civic life, and eroding the public’s trust in government. Put another way: Responsible development and use of AI/AS to further human values and ensure trustworthiness is the only kind that can lead to a sustainable ecosystem of innovation. It is the only kind that our society will tolerate.”