Jiahao Chen, PhD
Owner, Responsible Artificial Intelligence LLC
450 Lexington Ave, Unit 609
Manhattan, New York, 10163-9619
jiahao@responsibleai.tech

October 24, 2022

Department of Consumer and Worker Protection
New York City
42 Broadway
Manhattan, New York 10004-1617
Email: rulecomments@dcwp.nyc.gov

Dear committee members:

I am pleased to submit comments for *Requirement for Use of Automated Employment Decisionmaking Tools* (Ref. No. DCWP-21; "The Rules"). I am the owner of Responsible Artificial Intelligence LLC, a New York company offering algorithmic auditing and consultancy services. Previously, I was a Founder and Chief Technology Officer of Parity Technologies, Inc., a startup dedicated to modernizing model risk management and AI compliance; and also a Director of AI Research at JPMorgan Chase & Co., and a Senior Manager of Data Science at Capital One, where I started R&D teams for responsible AI and its applications to financial regulatory compliance such as the Equal Credit Opportunity Act; and also a Research Scientist at the Massachusetts Institute of Technology working on data science technologies. Since the start of 2022, I have spoken with multiple HR vendors, both startups and established companies, as well as a prominent legal firm, who have sought my input on how to establish compliance with the AEDT law ("The Law"). I would like to provide some comments on the real-world operational questions that have surfaced through my discussions with multiple data scientists, lawyers and vendors, as well as my thoughts on how best practices from federal regulators in employment and finance can be usefully translated into the context of the Law.

Vendor liability. One of the largest open questions around compliance for the Law is what liabilities vendors have. Most employers are unlikely to have in-house expertise to assess whether their use of such AEDTs are compliant, and thus will outsource compliance audits to third-party auditors. This is presumably the intent of the Law, to require such audits to be conducted. However, most employers also lack the ability to build their own AEDTs, and choose instead to purchase solutions from external vendors. In such situations, there is a separation of ownership between an employer's real-world employment data on one side, and the vendor's AEDT code and training data on the other. Auditors must thence navigate challenging internal politics to ensure both that the employer's data and the vendor AEDT is available for a successful audit, assuming that the vendor-client contract even permits such access.

Data ownership issues hinder testing of robustness. Modern data-intensive AEDTs that are built according to current best practices are not defined solely by their internal algorithmic rules, but also by the training data used to develop the AEDT, and also the development process of how models are built, selected against potential alternatives, and validated internally. It is therefore critical to assess the statistical similarity of the training data to the actual deployed use case. The performance of AI systems depends critically on the data fed into them - for example, most AI systems fed only men's profiles will not pick up on the absence of women, even if they were assessed to not have discriminatory outcomes when used on a gender-balanced test population. Therefore, the assessment of discriminatory outcomes must be evaluated in the context of the data for a specific population, which makes it difficult to purely absolve users of vendor solutions. Conversely, it is difficult for vendors to attest or certify that an AEDT will not produce discriminatory outcomes without making strong assumptions on typical usage of their solutions. In short, when vendors own training data while downstream users own test data, the two parties must engage in some form of data sharing to test for robustness and out-of-sample generalization error. Such data sharing must be done carefully to protect the identities in the data sets, particularly when it comes to EEO-1 demographic data and other personally identifiable information which could compromise privacy if leaked.

The risk of ethics-washing. There is an unavoidable conflict of interest that arises when companies pay for audits of their own products. Even in this nascent stage of development for the algorithmic auditing industry, controversy has already arisen over how some companies heavily censor the audits before release, or use the ostensibly neutral platform of academic publications to obtain validation for their reviews in the form of peer review. On one hand, it is important to recognize that compliance audits usually happen under attorney-client privilege, so that clients can address and remediate any negative findings without incriminating themselves in the process. On the other hand, the pay-to-play nature of auditing necessarily creates a conflict of interest that incentivizes keeping the auditee happy in exchange for future business and building relationships. Such concerns are of course not new, and have plagued financial auditing for decades. The experience of the financial services industry clearly points to the need for independent verification of audits, which are usually manifested in the form of regulatory audits by government entities.

Reproducibility requires data and algorithmic governance. The very act of auditing an AEDT implies that an auditor can independently reproduce claims made by developers about the properties of an AEDT, such as its lack of discriminatory impact or expected performance on a given test set. However, the act of reproducing an AI/ML system itself - to set up a replica of the production environment with a replica of a data stream that resembles real world conditions - can itself be a major engineering challenge. Successible reproduction of a production system in an audit environment that does not affect production data streams will be necessary to ensure that the auditing process does not inadvertently pollute and affect the use of the AEDT itself.

Data quality issues in demographic label collection. A related issue is that of data quality - the most highly predictive AEDT is still likely to fail if fed erroneous data. In the context of algorithmic auditing, data quality issues extend not just to the data fed into an AEDT, but also the EEO-1 demographic data and other personally identifying information that needs to be collected in order to correctly classify people by race, gender, age, and other protected classes. In practice, EEO-1 demographic data is voluntarily provided by applicants and employees, which means that people will voluntarily refuse to self-identify. Such refusal is not statistically random, but is disproportionately likely to occur when membership in a protected class, such as having a mental disability or being of a particular sexual orientation, carries social stigma or otherwise is likely to cause harm by "outing" someone to belong to some group. This missing not-at-random nature of demographic label quality ought to be considered whether or not discriminatory outcomes can be measured with sufficient statistical power, particularly if an imputation method like Bayesian Improved Surname Geocoding (BISG) is used to fill in missing demographic labels, as is commonly done in compliance testing for consumer financial services.

Construct validity. The need for AEDTs is greatest when there is an inherent scaling challenge to the number of decisions that have to be made. In the employment context, this usually shows up in the early stages of recruiting to narrow the funnel of applicants that are shortlisted for subsequent rounds of interviews. However, it is unclear if data collected at early stages of an employment decision, such as receiving a resume or video recording from a job candidate, will contain enough predictive signal to accurately predict a candidates suitability for hiring. In practice, AEDTs cannot predict something abstract like "employability", but instead compute metrics that purport to measure suitability scores or the like for such abstract concepts. An audit must necessarily assess the problem of construct validity, that a prediction target of an AEDT is indeed a valid and suitable quantification that operationalizes the employment decision being considered. Such considerations are of course of long-standing debate in federal employment laws; however, the algorithmic nature of decision-making and its use in making quantitative predictions bring such fundamental measurability concerns to the forefront of assessment. Many metrics purporting to quantify algorithmic bias implicitly assume that the prediction target of the AEDT is perfectly well-defined without any measurement ambiguity, which is unlikely to be true in practice. Therefore, the construct validity of the prediction target needs to be assessed critically to avoid false overprecision and overconfidence in the quantitative evidence for or against algorithmic bias.

The ethics of negative class sampling. A particularly thorny data quality problem goes by the name of reject inference in credit decisions, and is closely related to the problem of positive-unlabeled learning in other machine learning contexts. It is a problem for AEDTs that create a data asymmetry between positive and negative decision classes. For example, an employer incrementally collects more and more information about a candidate that passes multiple interview rounds. Conversely, a candidate not selected for an interview will have less data about them. This means that for hiring decisions, it is easier to assess false positives (a promising candidate that turned out to be a poor employee) than false negatives (a candidate that did not interview well that would have been a good employee). The counterfactual nature of the negative class makes assessments involving them difficult to assess in practice - someone removed from the candidate pool is by definition someone that was never placed in a job, and hence there was no real measurement of whether or not they were good at their job. A critical assessment of an AEDT's predictive value ought to include assessments of how well they classify the negative decision class, but if this class is not measured in any data set, then expert review is needed to validate negative decisions, or otherwise an experimentation framework is needed in order to test counterfactual changes to the AEDT prediction. There are obvious ethical risks to deliberately altering an employment decision for the sake of algorithmic assessment, as well as high costs of incorrect classification which will hinder the collection of real-world validation data. A well-designed audit should recognize the importance of negative class sampling, while at the same time have procedures in place to effect the necessary counterfactual testing without undue cost.

Intersectionality and subject privacy. The explicit call-out for intersectional testing across multiple protected classes is a welcome strengthening of current federal standards, which do not require testing of, say, race and gender, simultaneously. Nevertheless, intersectional concerns increase the risk of identification and hence loss of privacy for underrepresented groups. The more labels used to define a category, such as "people of gender A, race B, and age group C", the fewer people are likely to belong to that exact category. Taken to its logical extreme of testing every single protected class defined under federal employment laws, there is a risk that the intersectional categories are so fine-grained that only a single person may belong to that category. When such categories exist, summary statistics can leak information about a single person. In practice, the granularity of intersectional categories must be balanced against privacy concerns. I have some very preliminary research that indicates that differential privacy is a promising mechanism for achieving these goals in an algorithmic audit, although field testing will be required to validate the theoretical work we have been able to publish.

The fallacy of the four-fifths rule. The literature on algorithmic bias has unfortunately perpetuated a misconception of the significance of the four-fifths rule which the current rules are at risk of perpetuating and codifying. It is often claimed that the Equal Employment Opportunity Act enshrines the disparate impact ratio as the only legitimate metric for measuring employment discrimination, and that when it exceeds 80%, there is no finding of employment discrimination. In reality, tracing the historical development of the four-fifths rules reveals that it was only ever meant to be a bureaucratic rule of thumb for prioritizing cases for further regulatory scrutiny, and in fact the 80% threshold was effectively set arbitrarily in a 1971 meeting of the California Fair Employment Practice Commission as a compromise between a 70% camp and a 90% camp, a compromise that seems to not have been revisited with much scrutiny ever since. The arbitrariness of the four-fifths rule has been recognized by multiple federal courts in multiple court cases: courts have found that the 80% threshold is neither necessary nor sufficient to make a determination of discriminatory outcomes, and have admitted other forms of statistical testing, such as hypothesis testing for equality of means, in actual court cases. In short, the 80% threshold is arbitrary and fails to capture less severe discriminatory outcomes, particularly when the sample size is small and when the membership of people in protected classes is unclear.

To address these operational challenges, I would like to make the following recommendations for your consideration.

Recommendation 1. The City should invest in their own auditors and regulators to assess if audits need to be themselves independently audited, adapting relevant best practices from financial regulators and auditors where helpful.

Recommendation 2. The Rules would benefit from clarification on governance requirements for AEDTs and their associated data sets.

Recommendation 3. The Rules should clarify how robustness and generalization ought to be tested, and if so, how data sharing between different owners can be effected for the purposes of compliance audits.

Recommendation 4. The Rules would benefit from clarification on what liability vendors have for selling AEDTs to downstream clients, and to what extent (if any) these downstream procurers of AEDTs are able to shift liability to the vendor.

Recommendation 5. Regulators should work with standards-setting bodies, such as the National Institute for Standards and Technology (NIST), to develop and curate test data sets that represent typical populations which may be affected by AEDTs, so as to enable high quality testing of AEDTs that affords apples-to-apples comparisons.

Recommendation 6. The regulators should favor companies that have voluntarily adopted the NIST AI Risk Management Framework (RMF) or similar best practices for building and using AI systems. The regulators should issue more specific guidance aligned with the AI RMF to streamline compliance reviews.

Recommendation 7. The Rules should not codify any specific metric or threshold for passing or failing, but rather accommodate a possible plurality of valid metrics, and insist on tests of statistical validity rather than simply passing a numerical threshold.

In closing, I would like to congratulate the City for its innovation for enacting the Law, the first of its kind for the employment industry. The comments above are not meant to detract from the significance of the Law, but rather to highlight implementation risks that ought to be managed in order for the Law to have its desired effect to promote inclusivity and accessibility of job opportunities, improve transparency in high-stakes decision making, and reduce discimination in employment decisions. Please do not hesitate to reach out if I may be able to provide further clarifications on these comments.

Yours sincerely,

Jiahao Chen, Ph.D.