Thank you for the opportunity to comment on recent guidance for this landmark law. As vendors in the model audit space, we provide tools, technology, and services to employers and HR vendors who offer AEDTs to their clients. From this perspective, alongside our work with other regulating bodies, we hope to share some insight from the front lines of our work with clients, policy-makers, and academia alike.

## **Setting an (Insufficient) Standard**

As activists and academics in this space, our team has been working on issues of AI fairness, explainability, accountability, and model accuracy for many years. Our clients come to us because they know we possess deep knowledge of the many ways in which models can go wrong. In the past year, we have performed a number of audits prior to the release of this guidance that have sought to dig far deeper than the guidance requires. As has been previously stated by many in our community, the scope of the law and its proposed guidance, while desperately needed, may miss its intended effects of reducing discriminatory algorithmic behavior. Algorithms typically discriminate in a number of ways, from age to disability to race and gender, with many intersectional combinations thereof. By limiting the scope of the guidance as it currently stands, our prospective clients may be less motivated to pursue thorough examinations of their technology, leading to the shaping of our field’s form of practice. The remainder of our comments will seek to illuminate many of the complexities and conflicts within the practice of algorithmic auditing, in order to display a few of the ways in which the guidance may fail to uncover discrimination where it may continue to exist.

## **Vendors vs. Employers**

The proposed guidance contains a premise that may in many cases be untrue: that by modeling reporting from EEOC guidance, the commercial lift on industry will be reduced. However, employers are seldom the ones responsible for the development of AEDT tools, and employers far more frequently rely on vendors to provide the AEDT for a hiring process. This provides significant tension between the types of data needed for analysis, and the data that each entity possesses. Questions remain as to whether employers may themselves be allowed to simply repurpose the highest level reporting from the AEDT vendors, or whether these employers must also themselves perform an audit on their own limited candidate supply. One of the many ways that AEDTs can go wrong is due to generalizability errors. This is to say that vendor suitability may vary by employer (i.e. a tool that works well in healthcare may perform terribly and/or discriminate on engineers). The guidance, assuming it should take place at the global AEDT level, fails to address this kind of error, providing one of many ways that algorithms can be biased, but which will not be discovered in due course of a model audit seeking to replicate this draft for public reporting. One of the major conflicts between employer and vendor analysis for compliance with this law, deals with the question of whether and which entity may possess demographic data.

## **Demographic Data**

Because the guidance assumes it will be produced by employers, who may not have access to match scores at scale provided by their vendor, many of them have requested copies of this report from their vendors for their own websites. This presents a catch-22 for the vendors, in that they largely intentionally limit their collection practices for demographic candidate data, in order to comply with conflicting laws that require data minimization and candidate privacy like the GDPR. In the absence of self-reported data at application time, demographic data is difficult (and costly) to obtain. Candidates are often reluctant to provide this data post-hoc, as is well known in the financial industry who are permitted to collect this data only after applying to credit applications for compliance with fair lending laws, resulting in very low survey response rates. Even the simple act of retaining this data can generate enhanced risk for AEDT vendors, in that the leakage or theft of this data can result in fines or other legal penalties.

## **Availability of Intersectionality Supply**

We applaud the guidance in that it seeks to apply concepts of intersectionality in dividing the report into combinations of race and gender. However, when one requires that we investigate bias by multiple co occurring protected categories, we are making intersectional comparisons that split each of the broader categories into smaller ones, significantly creating large differences in who is represented in the audit. These more rare intersectional identities can often bear the majority of bias and prejudice throughout their careers and experience a great deal of gatekeeping. There have already been experiences wherein our investigations have resulted in intersectional slices for minority categories for which the amount of data in each category is so limited that the results are misleading, in that they do not represent statistically significant segments. Employers and AEDT vendors in this situation are faced with a significant problem, in that releasing the numbers for these categories may imply discrimination when the evidence of discrimination is simply not there. In some cases, there may even be categories for which no data exists. In this situation, we often see vendors turn to synthetic data or globally representative datasets which may not be sufficiently connected to the candidates truly subjected to the AEDT. The science on this notion is fairly limited, yielding great uncertainty in the audit community, and paving the way for insufficient audits that fail to represent reality.

## **Conflicting Laws**

Some good practices can conflict with each other. For example, data minimization and anonymization can protect our privacy, but also hide the very information we need to conduct an audit. Some industries (like the financial sector) use gender and race inference, or synthetic/global datasets to avoid the privacy issue, but this adds a thick layer of uncertainty to the audit that could cause some audits to under or overestimate bias, which we’ll elaborate on later.

## **Confounding Factors (i.e. Human Selection Bias)**

In defining “ground truth” it may be tempting to use some signal of approval (e.g. a candidate was hired, or a candidate was moved forward in the hiring process). However, these signals are human in nature, and therefore even more full of potential for discriminatory behavior (or the lack thereof) to obscure the behavior of the algorithm itself. An exceedingly DEI-consious hiring manager’s decision-making may cancel out a highly discriminatory algorithm, and of course the inverse is also true. A deeper examination of the tool’s training data, predictions, methods, structure, and behavior at scale in the context of the UX of the system can indeed illuminate this bias, but today’s guidance requires none of this type of reporting, and will therefore miss opportunities for improvement.

## **Mathification of Subjectivity**

When assessing algorithmic discrimination, it is vital to have a definition of “ground truth”. In the case of hiring, this notion is quite subjective, where the definition of “good candidate to job fit” can differ from organization to organization, and even among the hiring managers within that organization. This makes the challenge of a model audit an inconsistent one, where these definitions will vary significantly by audit vendor. In short, it is entirely possible to “game the system”, allowing vendors to provide audits that reflect a lack of bias where bias truly exists. The guidance in its current form does make way for one method to avoid assessing the human factor, by allowing for analysis of adverse impact by match score alone. However, later on in these comments we will detail just a few scenarios in which this simplified reporting may miss many forms of bias that remain, despite “passing” metrics. In order to assess algorithmic discrimination, a combination of quantitative and qualitative analysis is required, in order to contextualize and fully situate the impact to candidates amid the totality of the system. Candidate positioning, ranking, and display qualities matter a great deal to a candidate’s likelihood of receiving an offer. In addition, there are many standardization practices that AEDT vendors can undertake to limit discrimination that can only be uncovered through an assessment of their risk and control practices. By neglecting the qualitative elements of the field of algorithmic impact, the city paves the way for these reports to be misleading, and ultimately to fail to reflect real-world discrimination where it exists.

## **Demographic Inference**

As we’ve previously stated, employers may possess demographic data for their hired candidates, but the vendors who provide this technology often make active effort not to collect this vital information. As a result, these AEDT vendors often turn to methods like BISG to infer race and gender characteristics. BISG, as the most prevalent of these methods, was developed in healthcare research, and has been employed at great scale within the financial sector. However, besides concerns around accuracy, the methods themselves pose structural inequity. Race itself is a subjective attribute, and one which many have claimed can never be truly inferred. These methods also only allow for analysis on a gender binary, obscuring discrimination which may occur against others along the gender spectrum. An unintended consequence of this guidance may be the proliferation of these techniques, which have received deep scrutiny and criticism for their lack of inclusivity, and propensity for error. In fact, these error rates may in many cases be high enough to further obscure discrimination or lack thereof. If a set of candidates are improperly associated to the incorrect protected group, this may result in low enough accuracy to make the report incorrect, and therefore misleading. Additionally, common inference methods like BISG can only be effective in regions where we can assume that redlining, white flight, and gentrification have homogenized the racial makeup of the area. This seems broadly inadequate for a city as diverse as New York where there may be just as many Black John Smiths in the same zip code as there are white John Smiths. In our field, the vast consensus is that the only proper way to use demographic data in analysis is when it is volunteered from the candidates themselves. We recommend to our AEDT vendor clients that they engage in post-hoc surveys, despite our expectations that response rates will be low, because it will yield the greatest accuracy. These surveys take time, however, and in many cases the clients who have only begun this analysis in the second half of 2022 will not have adequate time to complete this initiative sufficiently prior to the release of their public reports.

## **AEDT Snake Oil**

At Parity, we are lucky enough to have the principle and luxury of refusing to perform audits for technology that entirely fails to accurately represent accurate matching capabilities to employers. We have been approached by AEDT vendors who seek to provide technology that attempts to apply pseudo-scientific principles to the concept of job fit, and these vendors may be able to display reports that make them seem somehow fairer or more accurate than tools that avoid these forms of algorithmic phrenology. Stated another way, a tool that is always incorrectly guessing a candidate’s fit for the job, or that may approve or deny every group equally along vectors of race and gender, might appear fair. However, this fairness is only trivial – an algorithm that denies everyone for every job may be fair, but it is also not useful. These pseudo-scientific algorithms present far greater danger for candidates with disabilities, but the law in its current form will fail to capture this discrimination entirely.

## **Inaccruate Reporting of Discrimination/Lack of Discrimination**

Finally, that this guidance is so high-level allows for many opportunities to obscure discrimination, or to report discrimination where it does not exist, leading to undeserved risk and scrutiny to the vendors/employers. There are many scenarios under which discrimination may be obscured, but to name a few:

1. Imagine an algorithm that performs quite fairly on nursing jobs, but discriminates to a great degree in engineering. If these numbers are high enough in some categories to cancel out the low scores from another category, then the reports will appear to “pass” the 4/5ths rule, but discrimination will remain.
2. Imagine a situation where an employer receives a set of applicants for a position wherein all of the female candidates for a position are, in fact, highly qualified for the job, and all of the male candidates are similarly unqualified. When one set of match scores are high and the other set is low, it can appear that adverse impact exists in favor of women against men, when the resulting metric may fail to represent a lack of discriminatory bias in the tool itself, but instead a feature of the applicant base. This is a direct result of the subjectivity of quantifying “job fit” in mathematical terms.
3. It may be the case that, due to systemic inequity and historical adversity, some intersectional slices of demographic pools may receive lower match scores than others. This may not be the result of a discriminatory tool, but instead a feature of the applicant population as it exists today. Correcting for bias under these circumstances is recommended by academia, but itself may pose a form of “disparate treatment” by virtue of adding weight or altering thresholds to cater to one disadvantaged group over the other.
4. Due to the lack of demographic data, some categories may have insufficient amounts of representation to be accurately quantified, leading to numbers that skew inaccurately in a way that would not reflect discrimination that exists at scale.
5. When demographic inference is employed, the error rates may be so high as to make the resulting metrics adequately inaccurate such that they will not reflect reality, be that in the form of discrimination or a lack thereof.
6. Imagine a tool that is simply of very poor quality. This tool may be trivially fair, in that it approves or denies all candidates equally because it simply does not work. Employers choosing vendors may be misled into thinking that the tool is worthy of use by virtue of this needed reporting, when in fact it simply “stabs in the dark” at job fit, and may present cases of individual discrimination, especially with regard to demographic categories not represented by the guidance as it stands today.

These may seem like toy examples, but from our work with clients and in research, we find them to be fairly common. Today’s guidance may miss these situations, but the field of algorithmic interrogation has provided myriad tools and methods to uncover these scenarios, and we would encourage the city to pursue guidance that is more closely in line with the latest our field has to offer.

## **Conclusion, Recommendations, Next Steps**

We’d like to reiterate our gratitude for the opportunity to provide comment. Significant questions remain on the scope of the law that would preferably be answered in advance of the compliance deadline:

1. Is a broad, global vendor analysis sufficient to each employer who uses the AEDT? Or should each employer tailor the report to the candidates they’ve assessed?
2. Will race/gender inference suffice for the analysis despite its possibility for decreased accuracy, and if not, what methods do you recommend when demographic information is unavailable?
3. When intersectional slices for rarer combinations of categories are present, and the amounts may not be statistically significant themselves, what sort of reporting does the city recommend?
4. Would universal or synthetic datasets suffice for the analysis, even if these datasets may not be representative of the candidates truly subjected to the system’s decisions?
5. Are other forms of screening models (e.g. “culture fit, engagement probability, geographic closeness”, etc.) within the scope of the law? Or is the scope limited to assessments of job-to-candidate fit?
6. Will the city consider some form of vendor certification moving forward in order to limit the ability for tools and employers to game the system by choosing unscrupulous providers?
7. Our field of algorithmic scrutiny is rapidly advancing, will the guidance make room for the advancements not currently included in the guidance, and continue to evolve with the pace of science?
8. Will the city consider extending the compliance deadline in order to provide more time to employers and vendors to begin the arduous practice of collecting demographic information from candidates?

We would be happy to further engage with DCWP in order to clarify these questions or to improve guidance for this upcoming or future years, and look forward to your feedback.

Thank you,

The Parity Team