DATA&
SOCIETY

January 23, 2023

To the New York City Department of Consumer and Worker Protection Commissioners:

Thank you for the opportunity to comment on the proposed rules for implementation and enforcement of Local Law 144, regarding the use of automated employment decision tools (AEDTs) in hiring and promotion processes.

Data & Society is an independent, nonprofit research institute studying the social implications of data-centric technologies and automation. We produce empirical research that challenges the power asymmetries created and amplified by technology in society, and work to help ensure that artificial intelligence (AI) systems are accountable to the communities within which they are applied.

Local Law 144 is one of the first laws in the world to mandate an independent audit of any algorithmic systems for bias, and therefore this law and rule-making process has important implications beyond the jurisdiction of the DCWP. Not only are such systems used in employment contexts, they are increasingly used across the economy and government in sensitive domains, such as distribution of social welfare, educational opportunity, housing, and access to financial resources. As is well-documented in scholarly literature, government reports, and investigative journalism, the use of machine learning to train these computational systems is prone to bias against vulnerable and historically disadvantaged groups of people. At their core, these systems learn to replicate the past decisions and behaviors recorded in their training data—these systems predict how we *would have* acted in similar contexts, leaving little room for adjusting how we *should have* acted. Regardless of the efficiency that machine learning systems promise to those who use them, *society has an obligation to ensure that such efficiency is not gained on the backs of vulnerable populations*.

Local Law 144 addresses that obligation by requiring those who deploy these systems in an employment context to transparently account for how their systems behave toward the actual population of job seekers, notifying the public and applicants of their use, and giving applicants the right to request alternative methods. The City of New York is right to pry open this black box for job seekers, and should continue to do so for other domains in the future.

However, as an organization of scholars and policy experts in the social consequences of data technologies, we are concerned that some details of these proposed rules will dramatically blunt the effectiveness of this law and subvert the intent of the New York City Council. We note that these rules may unnecessarily limit the scope of these auditing obligations in three ways:

1. **Narrowly defining Automated Employment Decision Tools;**
2. **Misunderstanding how machine learning bias is propagated; and**
3. **Restricting bias audits to gender and race/ethnicity.**

Defining Automated Employment Decision Tools

The proposed rules define automated decision tools in employment to mean a system substantially assists or replaces discretionary decision making:

i. to rely solely on a simplified output (score, tag, classification, ranking, etc.), with no other factors considered;
ii. to use a simplified output as one of a set of criteria where the simplified output is weighted more than any other criterion in the set; or
iii. to use a simplified output to overrule conclusions derived from other factors including human decision-making.

We are concerned that this definition is so narrow as to exclude the majority of AEDT applications on the market, and misses the core motivations behind LL144.

This definition appears to assume that the biased outcomes that result from AEDTs derive only from *automated* decisions. However, the economic rationale of AEDTs for most employers is not to render a final hiring or promotion decision on the basis of machine learning outputs alone, and the market for such a tool therefore has few (if any) options that operate in a fully automated fashion. Some very large employers, most notably Amazon warehouses, utilize what appear to be fully-automated recruitment software that is developed in-house (though there are few public accounts of how these hiring processes actually work). While those systems certainly deserve scrutiny, and Amazon's many potential warehouse employees deserve the protections offered by LL144, it does not appear that the City Council intended to limit the scope of the law to a very small number of large employers.

Rather, most employers use these systems to create time and economic efficiencies in decision-making by the humans tasked with final hiring or promotion decisions. The practical reality in most cases is that humans still make the final decision from a pool that has been filtered, sorted, scored, and/or narrowed by prior computation. However, this definition restricts

the scope of the rules to systems where an algorithmically-generated score is either the sole or predominant factor in the hiring or promotion decision.

Our strong suspicion is that very few employers who one could judiciously say currently deploy an AEDT have a system that would meet reasonable interpretations of this proposed definition. *Most job seekers who are algorithmically scored would not be protected under these rules, which we do not believe was the intent of the City Council in passing LL144.*

At the very least, this proposed definition leaves open the door to subsequent legal challenges that may gut the intent of LL144. In particular, the language in subphrase (ii) raises the question of how stakeholders might measure "weighted more than any other factor"—the ambiguity here leaves open specious legal interpretations that nearly every employer could adopt to avoid conducting audits. Does "weighted more" mean that if the algorithmic score is weighted 49% then it is not open to the scrutiny of an audit? Does it means that if the automated score is weighted at 20% and eight other factors are weighted individually at 10% and collectively 80%, then the system is subject to the rules? How would the DCWP ask employers to reliably account for the weights that are used?

Furthermore, there are abundant studies examining the highly complex and fraught relationship between human discretion and algorithmic scores. The story that emerges is that a multitude of factors determine to what extent humans utilize discretion when presented with algorithmic scores. Even when prompted to use discretion, humans in organizational contexts where they are otherwise incentivized to trust the computer will follow the algorithmic predictions the vast majority of the time. Given the many different corporate structures, incentives, and internal information systems at private employers over which DCWP has no insight or control, it would seem that the spirit of LL144 requires assuming that *any* system which uses algorithmic scores is potentially a source of algorithmic bias and therefore subject to audit.

We also note that the only change in the wording of this definition between the prior proposed rules (considered for public comment in October, 2022) is removal of the word "modify" from subphrase (iii), replacing "overrule or modify conclusions" with simply "overrule conclusions." This change again significantly reduces the number of AEDTs that would fall under scope, reducing the practical reach of these rules.

We suggest to the DCWP that the definition of AEDTs used in the rule-making process should hew more closely to that plainly stated in the text of LL144:

> "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." [emphasis added]

The definition offered in this proposal accounts in practice only for those systems that *replace* discretionary decisions, not for those that *assist* discretionary decisions, and therefore is too dependent on the methods used by the developer and/or the intent of the employer. The simplest and most direct route to providing the protections of job seekers that LL144 plainly intends is to *subject all algorithmic scoring systems to independent audits*. Audit obligations should apply without consideration of the degree to which the employer *intends* to weight those scores.

Misunderstanding of how bias is propagated in machine learning systems

The second significant flaw in these proposed rules is that the definition of "Machine learning, statistical modeling, data analytics, or artificial intelligence" is overly narrow, and as a consequence misunderstands how bias operates in machine learning. Potential routes to biased hiring and promotion decisions could be technically excluded.

The proposed definition is as follows:

> Machine learning, statistical modeling, data analytics, or artificial intelligence. "Machine learning, statistical modeling, data analytics, or artificial intelligence" means a group of mathematical, computer based techniques:
>
> i. that generate a prediction, meaning an expected outcome for an observation, such as an assessment of a candidate's fit or likelihood of success, or that generate a classification, meaning an assignment of an observation to a group, such as categorizations based on skill sets or aptitude; and
> ii. for which a computer at least in part identifies the inputs, the relative importance placed on those inputs, and other parameters for the models in order to improve the accuracy of the prediction or classification; and
> iii. for which the inputs and parameters are refined through cross-validation or by using training and testing data.

The specific error here is found in (ii): "for which a computer at least in part identifies the inputs." In machine learning, algorithms are used to find patterns in a collection of features (categories of data, such as educational level, degree, years of experience, previous job titles, previous employers, etc.) that statistically indicate a certain outcome is likely to occur (such as a candidate is likely to be successful in a job role). The pattern that indicates success is then structured as a *model*, a set of mathematical instructions for an application to predict future outcomes based on live inputs (such as the content of new applicant resumes). The efficiency of

machine learning is finding the *optimal* arrangements (weights) of features to predict success in the objective function (such as finding a good candidate).

In *some* cases of machine learning, the computer chooses which features/inputs to utilize in building the model. Those methods are often known as "deep learning" wherein a very large unstructured dataset of features—many of which may be facially irrelevant in human judgment—is analyzed by the algorithms to generate an optimized model. Data scientists are rightfully concerned about how deep learning can unintentionally and inscrutably propagate historical biases embedded in their training data. However, deep learning is only likely to comprise a small proportion of the types of machine learning utilized to construct AEDTs because the data available in a resume is already highly structured, labeled, and pre-determined by the expectations of job-seekers and hiring managers. Before machine learning has entered the picture the inputs are already chosen simply because resumes are largely standardized, which means that many AEDTs could be technically excluded by this definition.

Additionally, developers of machine learning systems are often substantially engaged in crafting the models—despite the marketing rhetoric around automation, there is nearly always significant human input and artfulness that goes into shaping these applications and services. It is likely a rare occurrence for the computer to choose relevant features, optimal weights, and parameters alone. The developer's choice of statistical techniques may also introduce opportunities for bias such that even the most rudimentary forms of machine learning result in bias. Similarly, these systems are often customizable by the employer. A hiring manager may manually choose certain weights (defined here as "the relative importance placed on those inputs") that still drive algorithmically-biased consequences. In other words, especially in machine learning systems meant to intervene in human social processes like AEDTs do, human discretion in the construction of the model is just as likely to introduce bias as deep learning techniques.

Therefore, it is possible that AEDTs which use fairly simple (and very common) machine learning techniques to evaluate candidates on the basis of their resumes would evade this definition.

Of course, some ambitious AEDTs may use features/inputs that require deep learning methods, such as intelligence tests, personality tests, or biometrics. Such applications may require the machine learning system to model the relevant features at a fine granularity, such as the pattern of mouse movement to complete a task or the structure of a person's face when smiling in a video interview. However, even in those applications, at a gross scale humans are still choosing the relevant inputs, such as efficiency and emotional state. Using the rules as currently proposed leaves the DCWP open to legalistic objections and evasions on this question.

In short, only the actual measurement of bias really matters here—exactly how the system is constructed is largely irrelevant to the question of whether bias may be present.

Simply striking point (ii) in this definition would resolve this error and still leave DCWP with a defensible and adequately capacious definition. Alternatively, the conjunction "and" could be replaced by the disjunction "or" in (ii) to clarify that *any* of those techniques is classified as machine learning.

Restricting bias testing to race and gender

The proposed rules only require bias auditing of gender and race/ethnicity features as defined by the US Federal Equal Opportunity Commission (EEOC). There is good reason to use these standardized categories from EEOC rules, insofar as they are commonly understood, nearly universally collected, and do not generate conflict with other statutes. We affirm that the DCWP is correct to use these categories in the bias audits.

However, we note that LL144 does not specify any requirement to limit bias audits to race and gender features, and therefore leaves the door open to auditing against a more expansive list. The EEOC categories should be a floor, not a ceiling; all AEDTs should be audited for bias along these features, but other biases should be audited for if the system implicates relevant features.

For example, multiple commercial AEDT products analyze audiovisual content of video interviews to predict personal characteristics such as personality or affect. In one audit[1], a group of journalists found that a commercially-available personality profiling AEDT generated significantly different results if candidates wore glasses, changed their background to include a bookshelf, put on a headscarf, or changed their lighting conditions. Obviously, none of these characteristics are correlated with stable personality features predictive of job performance, and thus the product is itself dubious. However, such products can also introduce unexpected biases: affect can be associated with gender and sexual identity, headwear can be correlated with religion, and eyeglasses are a prosthetic to correct for a disability that doesn't affect job performance. Similarly, text-based tests or cognitive tests may be swayed by neurodivergence or cognitive disability unrelated to job performance and/or amenable to reasonable accommodation as required by the ADA. None of these known, well-demonstrated, and illegal types of bias in AEDTs would be accounted for in the proposed rules.

Our recommendation to the DCWP is that AEDTs should be audited according to the type of bias they are likely to propagate based on the inputs chosen by the developers and deployers. Developers and deployers of AEDTs are responsible for choosing the features used in these

[1] https://interaktiv.br.de/ki-bewerbung/en/

systems. Multiple ethical algorithm design resources for tracking these risks are publicly-available, including resources developed by the National Institute of Science and Technology. The independent auditors mandated by LL144 are capable of identifying such features and the bias risks associated with their use, and responsible AEDT developers already do so. A more expansive audit would not pose an undue burden.

Beyond transparently accounting for bias, this would also promote the desirable consequence of weeding out "algorithmic snake oil" offerings in the AEDT marketplace. Algorithmic systems will always excel at making measurements and offering predictions in a manner that appears useful and economically valuable. But whether those predictions are relevant to the objective function (e.g., job performance), or desirable by society at large (e.g., fair opportunity), is often unanswered. Algorithmic snake oil is a common mode for injecting unfairness into a system because irrelevant measurements can be disguised as objective mathematical judgment.

But there is a simple solution to this problem: if the consequences of including a particular feature cannot be included in a bias audit, then that feature need not be used. Transparent and independent bias audits are one mechanism to force a developer to justify their choice to include certain features and prove that it does not create illegal bias, *but only if the relevant types of bias are accounted for in the audit.*

There is no justifiable reason to treat the EEOC gender and race/ethnicity categories as a ceiling rather than a floor and permanently exclude other, known types of bias. If the audits are restricted to EEOC gender and race/ethnicity categories for now, there should be explicitly stated intent to include more categories in the future as audit methodologies and data collection are improved over time and become routine.

Conclusion

We believe that LL144 is a much-needed and ground-breaking intervention in a market that has been harmful and largely neglected by regulators. However, the proposed rules are too narrow to provide the protections intended by the City Council.

The DCWP should pursue a simple principle in the next round of rule-making: developers and deployers of AEDTs are responsible to measure and transparently report how their systems behave when imposed upon the job-seeking public regardless of how the systems were constructed. This version relies too heavily on the presumption that bias is introduced *by* machines that replace human judgment, when in fact *all algorithmic bias is introduced and/or mitigated only by the choices of the developer and deployer* to engage machine learning for these tasks.

When those choices result in bias, they alone are accountable in each case.

Thank you for the opportunity to include our remarks on this critically important public policy. We hope that the work DPWC is doing on this topic can shape other efforts in the future.

Sincerely,

Jacob Metcalf, PhD
AI on the Ground Initiative, Program Director