SIOP White Paper Series

SOCIETY for INDUSTRIAL and ORGANIZATIONAL PSYCHOLOGY

SCIENCE FOR A SMARTER WORKPLACE

Algorithmic Justice

Georgi P. Yankov, Breanna Wexler, Sarah Haidar, Sukesh Kumar, Jimmy Zheng, & Ann Li
Development Dimensions International (DDI)

A White Paper prepared by the HR/Business Subcommittee of the Visibility Committee of the Society for Industrial and Organizational Psychology, 440 E Poe Rd, Suite 101 Bowling Green, OH 43402

Table of Contents

Authors 1
Introduction 2
Defining Algorithmic Justice 3
Research Guidelines for Algorithmic Justice 4
Discovering Algorithmic Bias 7
Legal Implication Related to the Use of Algorithms 8
Conclusion 11
References 12

Authors

Georgi Yankov, PhD, is a research scientist at DDI's world headquarters in Pittsburgh. His expertise is in psychometrics and individual differences. At DDI, he designs and validates assessments for leadership selection and development. In recent years, Georgi has been working on integrating machine learning in the measurement of candidates' personality and behavior.

Bre Wexler, PhD, JD, is a consultant at DDI's world headquarters in Pittsburgh. She works on reporting and analytics projects for DDI's clients using leadership testing and assessment data, and leads several internal reporting initiatives. Bre is also responsible for creating and maintaining the surveys and reports that demonstrate the impact of DDI's assessment and development solutions.

Sarah Haidar, PhD, is a product development consultant at DDI's world headquarters in Pittsburgh. She creates and optimizes products that help select and develop leaders. Recently Sarah has been exploring with the application of advanced technologies such as virtual reality and natural language processing in talent assessment and development.

Sukesh Kumar, MSc, is a data science engineer at DDI's world headquarters in Pittsburgh. His primary work includes driving insights by implementing machine learning algorithms for software product development. Recently, he was responsible for applying natural language processing algorithms to create virtual personal assistants, adaptive web-based training modules, and language understanding metrics in virtual reality.

Jimmy Zheng, MA, is an intern in DDI's product development group.

Ann Li, PhD, is an intern in DDI's Survey, Testing, Assessment, and Team Services (STATS) group.

Introduction

With big data availability and advances in computational power and statistical methods, the reliance on advanced assessment techniques such as machine learning (ML) algorithms will continue to grow. Despite the obvious value of such technologies (e.g. faster and more engaging selection processes, potentially better prediction, personalized recommendations, etc.), there is evidence that they could perpetuate biases and thus may violate workplace antidiscrimination laws. This white paper is an effort to define the problem under the concept of algorithmic justice, as well as to provide practical information and advice for the use of ML algorithms in workplace assessment. The white paper proceeds in four steps by:

1. defining the emerging concept and scope of algorithmic justice,
2. outlining the research guidelines for ML algorithms used in workplace assessment,
3. explaining why ML algorithms can be biased, and
4. discussing potential legal threats when using algorithms in workplace assessment.

The term algorithmic justice was coined by Joy Buolamwini, who also founded the Algorithmic Justice League at MIT to move toward equitable and accountable use of artificial intelligence (AI). Specifically, algorithmic justice is evolving from practical concerns with the accuracy of ML algorithms to classify individuals from protected classes. The widespread use of ML algorithms compounds this concern. ML algorithms are being used in various contexts such as medical diagnosis (e.g., using facial cues as predictors of health indicators such as fat percentage, body mass index, and blood pressure; Stephen, et al., 2017), law enforcement (e.g., using the Next Generation Identification-Interstate Photo System to help solve crimes), and employment.

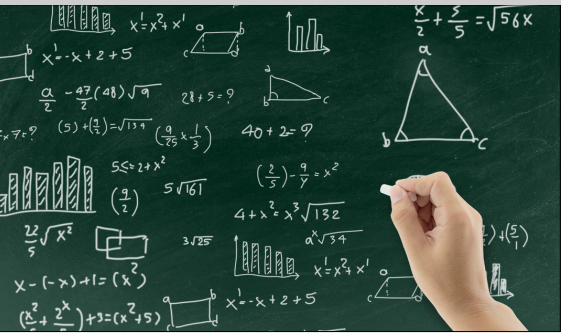
Currently, ML algorithms' key promise for faster and accurate prediction of human characteristics is more desired than accomplished. First, a controversial use of ML algorithms is for facial recognition. For example, Buolamwini and Gebru (2018) found that these algorithms can discriminate based on race and gender. A potential reason for these differences in classification accuracy is that when facial images are captured in non-ideal, everyday life conditions, algorithms cannot accurately detect the facial features (e.g., jaw drop, blink) that human coders can detect (Barrett et al., 2019). Second, ML algorithms have also been credited with predicting personality. However, this is an overstatement given overall research findings. For example, Youyou et al.'s (2015) study of digital footprints found that when fed with only 300 Facebook likes, Extraversion scores produced by modified linear regression related more strongly with one's self-reported Extraversion than the Extraversion scores provided by one's spouse. At the same time, Nguyen et al. (2014) aimed to predict personality based on the automatic extraction of nonverbal cues but were successful only for Extraversion. More recently, Escalante et al. (2018) found positive judgement biases toward female subjects on all personality factors except for Agreeableness as well as negative biases toward African American subjects.

To underscore the potential issues with the widespread use of ML algorithms today, the following table presents prominent examples of algorithms gone wrong or right, as well as their practical outcomes and societal implications. Thus, it is not surprising that members of the US Congress raised concerns to the Federal Trade Commission, Federal Bureau of Investigation, and Equal Employment Opportunity Commission regarding the potential harm associated with the application of such technologies . Senators have also proposed a bill on algorithmic accountability.

Company	Algorithm	Outcome	Social implications
Amazon*	Resumé reviewer: Trained the algorithm by observing patterns in resumes submitted to the company over a 10-year period. Most resumés were from men.	AI taught itself that male applicants were preferable rewarding resumes that included words such as *executed* and *captured*, which are words more often used by men, and penalized resumes that included words such as *women*.	Could widen the gender gap in an already male dominated industry.
District of Columbia Public Schools**	The evaluation system IMPACT rates teacher's performance primarily on classroom observations and student test scores. Teacher performance ratings determine compensation and job security. The small number of students per teacher makes estimating student test scores statistically unsound.	IMPACT has resulted in the firing of many educators, placed hundreds more on notice, and left the rest frustrated and scared about their job security.	The influence of IMPACT is felt the most by teachers in DC's poorest school districts where minority teachers tend to reside. It perpetuates workforce inequalities and exacerbates an already alarming shortage of teachers of color.
Starbucks/Kronos***	Global labor scheduling system uses data on weather patterns, sales, and customer foot traffic to predict labor demand and schedule employees more efficiently. More employees were scheduled during busy times and less were scheduled during slow times.	Workers were given irregular schedules; sometimes having to close the store at 11pm and return at 4am to open it. They also did not receive adequate notice of their new schedule (less than a week's notice)	A single mother trying to work her way through college while working at Starbucks has to put school on hold due to irregular work hours and lack of scheduling notice.
Textio****	Uses NLP on job listings, recruiting emails, and hiring results to identify language that may deter applicants of minority from applying.	Helps make job descriptions more gender neutral. Attracts a more diverse group of candidates.	Increases the diversity of the applicant pool and subsequent hires.
Blendoor*****	Removes data from resumes that can result in algorithmic bias (e.g., name, address).	Helps remove conscious and unconscious biases from hiring.	Increases the diversity of the applicant pool and subsequent hires.

* https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html
** https://www.air.org/edsector-archives/publications/inside-impact-d-c-s-model-teacher-evaluation-system
*** https://searchhrsoftware.techtarget.com/news/4500252451/Kronos-shift-scheduling-software-a-grind-for-Starbucks-worker
**** https://www.upturn.org/reports/2018/hiring-algorithms/
*****https://medium.com/allraise/stephanie-lampkin-founder-ceo-of-blendoor-all-raises-enterprise-saas-womancrush-wednesday-5b6990570a28

Nowadays, algorithms are no longer used solely for mathematical computations but are tasked to solve complex problems such as predicting job performance.



Defining Algorithmic Justice

All ML algorithms follow predefined steps and aim to automatically detect patterns from data. Nowadays, algorithms are no longer used solely for mathematical computations but are tasked to solve complex problems such as predicting job performance. Each ML algorithm follows a different "statistical recipe steps" for summarizing and discovering relationships in data. In general, algorithms should not evoke fear because they have existed for centuries; for example, a knitting pattern is as much an algorithm as any of the natural language processing (NLP) ML algorithms. Developing algorithms for the assessment of humans is a ground over which neither computer science, math, engineering, nor psychology can solely claim ownership. The need for multidisciplinary collaboration among psychologists, computer scientists, and engineers is also overdue. The real need is for transdisciplinary solutions (Liem et al., 2018) so that algorithms do not feed us with unreliable and potentially illegal recommendations.

Industrial-organizational psychology, however, might be useful for defining what is a just or fair ML algorithm for workplace assessment. This is because research emanating from I-O psychology relates to similar topics such as the use of structured assessments for employment decision making. For example, Gilliland (1993) proposed an organizational justice model for selection procedures. This model, as applied to fair ML algorithms, can be readily summarized in terms of the three types of justice that I-O psychologists recognize:

Distributive justice:

- Does the algorithm provide applicants with an assessment decision (note that this is not necessarily a hiring decision) commensurate with their knowledge, skills, abilities, and other characteristics (KSAOs)?
- How actual outcomes from the algorithm applications lead to fair and equitable outcomes (e.g. hiring decisions) for those assessed?

Procedural justice:

- Does the algorithm assess job-related candidate characteristics (e.g., attention to detail) versus random data on the candidate scraped from social media (i.e., data that do not give the candidate the opportunity to perform well for the purpose of the assessment)?
- Were these characteristics assessed accurately (i.e., with low measurement error) and thoroughly (i.e., not assessing attention to detail with a score from a 1-minute game)?
- Was the candidate able to object to the algorithm's results?
- Was the candidate able to ask questions and prepare their performance to present themselves to the best of their abilities? For example, in automatically scored video interviews, was the candidate able to ask a question during the interview which allowed for a customized answer that helped the candidate during the rest of the interview?
- Did the algorithm assess candidates consistently? That is, did it use as predictors the same personal characteristics (e.g., the same personality scores) for every candidate?

Interpersonal justice:

- If the candidate was rejected, did the algorithm user give honest feedback to the candidate? Was the candidate told which personal characteristics and assessed variables were analyzed by the algorithm and how they led to the assessment decision?

If the answer to each of these questions is yes, then the algorithm is more likely to be perceived as a fair/just algorithm.

Research Guidelines for Algorithmic Justice

This section interprets the *Principles for the Validation and Use of Personnel Selection Procedures* (Principles; SIOP, 2019) and their research guidelines for ML algorithms applied in workplace assessment. Below we discuss five primary guidelines as applied to ML algorithms.

First, ML algorithms used to harvest candidate data from social media for professionals (e.g., LinkedIn) and resumés are considered modern selection procedures (p. 4). Just like for established selection methods (e.g., interviews, cognitive tests), the algorithm user must provide inferential evidence about the job relatedness of the algorithm scores. Noteworthy, it is not the reliability of the ML algorithm that is evaluated per se but its validity. A valid ML algorithm is characterized by substantive theory and evidence supporting the inferences and interpretations for future job performance derived from the algorithmically produced scores (p. 5). Therefore, the algorithm provider should not exclusively present evidence of the accuracy (i.e., reliability) of the algorithm and automatically infer that the algorithm is valid for predicting job-related outcomes.

Second, the *Principles* require that the provider should specify the variables (called predictors by psychologists and features by computer scientists) that the testing procedure will measure (p. 20). Although it might be difficult to trace how the predictors and their combinations (like the "black box" neural layers in NLP) relate to the outcome in question (i.e., predicted future job performance), the *Principles* require such analysis be performed individually for each predictor (p. 21-22). Specifically, algorithmically produced selection scores need to be justified not only methodologically but also conceptually with regard to their linkage to the criterion/outcome variable (p. 22). Generally, scoring and transformations performed by the algorithm should be described as fully as possible (p. 63). Even if not published due to copyright, the algorithm's computational model and validation should be documented. If the algorithm is contested legally, this document should be available so that an independent and statistically savvy evaluator of the selection procedure can reproduce the algorithm and its scores.

Third, because ML algorithms combine multiple (sometimes thousands) predictors/features, algorithm developers and users need to consider and document how the algorithm combines these predictors (p. 25). Given the statistically complicated nature of some ML methods, assigning weights to each predictor as in multiple

Even if not published due to copyright, the algorithm's computational model and validation should be documented.

regression is unrealistic. This, however, does not mean the algorithm providers do not need to explain and document how the combination is performed and how types of predictors interrelate (e.g., correlations and covariances). For example, within the algorithm's calculations, how do the features derived from facial recognition relate to features derived from the sentiment analysis of the candidates' text? Furthermore, these interrelations should not be only justified from one research sample that is (often artificially) split into training and test samples to develop the algorithm, the inter-relationships must be cross validated in another sample (p. 26) or at least with multiple randomized training and test samples (Friedler et al., 2019). Finally, all samples used to train, test, and cross-validate the algorithm must be described in terms of demographic composition, population representation, and potential range restriction (p. 63).

Fourth, regarding the most used selection method—interviews—ML algorithm users have to be aware that selecting some candidates with a traditional human-proctored interview and some candidates with an algorithm-scored interview is dangerous. Even if the latter is accurate and valid, candidates might be receiving different assessments whose scores might not be equivalent as the candidates' experiences during both interview types are different and the assessed behaviors are no longer equivalent (p. 48).

Fifth, regarding cutoff scores, ML algorithm users evaluating candidates against a cutoff score should provide evidence if the predictors/features relate linearly to the outcome. That is, if a higher score on each predictor relates to higher predicted future job performance. At the same time, it is advisable that users of algorithms document their selection goals and circumstances that led to employing an algorithm with a cutoff score versus an algorithm that ranks candidates (p. 58).

Based on the five clusters of the challenges above, we provide a checklist for evaluating the research rigor of ML algorithms integrated in assessment tools. HR officers can ask algorithm providers these questions to be more confident in the validity of the assessment tool of interest.

Checklist for Evaluating ML Algorithms

- Does the ML algorithm's training data represent individuals from protected classes and do the features convey protected data?
- Did the developers of the ML algorithm make clear that their input training data follows the population distribution of future input data? The same question applies to the outcome and measurement error of predictors and outcomes.
- Is the algorithm trained on supervisory ratings of job performance? Algorithms might have learned the biases of hiring managers through the training data used to calibrate the algorithms.
- Who designs the ML algorithm matters—what was the composition of the team that created the algo-

ML algorithm users have to be aware that selecting some candidates with a traditional human-proctored interview and some candidates with an algorithm-scored interview is dangerous.

rithm? Were the computer scientists informed by measurement and legal experts?

- Do the assessment scores produced by the ML algorithm conform to established psychological theory? If in psychological theory the predictors relate to the predicted outcome linearly (e.g. conscientiousness predicts job performance), then the ML algorithm scores must reproduce this expected relationship.
- Does the ML algorithm use human judges to define the features to be extracted from the raw data? Defining the features is like defining the predictors in psychological assessment and is subject to proof of job relatedness. Decisions on creating features directly affect whether an algorithm might turn out to be fair or not.
- Does the algorithm use substitute variables or data (called proxies) for variables because the intended variables were unavailable? If so, are these proxies evaluated for systematic bias? For example, Volpone and colleagues (2015) found that using job candidates' credit scores disproportionately disadvantaged individuals of color. Instead, more race-neutral variables were recommended to be used.
- Was the ML algorithm continuously improved? For example, in a selection setting, if applicant responses are transcribed from voice to text and then analyzed using an algorithm, it would be important to evaluate intergroup scores to assess whether there are systematic differences due to different accents.

Decisions on creating features directly affect whether an algorithm might turn out to be fair or not.

Discovering Algorithmic Bias

This section sheds light on why and how algorithms become biased. We have included it with the awareness that it might be a little complex depending on the professional backgrounds and prior knowledge of some of our readers. Nevertheless, we believe that in is in the interest of these readers to know more on discovering algorithmic bias even if they do not fully grasp the intricacies. For example, HR practitioners can use information in this section to come up with talking points and targeted questions for vendors of ML algorithms.

ML algorithms try to learn patterns in the data. In other words, they try to approximate a function that defines a relationship between input variables (features) and the output label that we want to predict. In this process, assumptions are made that lead to algorithmic bias. Assumptions depend on each algorithm, but a few examples include when input variables are assumed to be independent of each other, each of them is assumed to be identically distributed under unknown probability distribution, a linear classifier assumes that the decision boundaries are linear, and so on. However, without assumptions the algorithm would have no better performance than a random guess, a principle of "No Free Lunch Theorem."

There could be numerous reasons for having biased data. It could be due to over/undersampling, outdated data, a distribution of a feature that is not representative of the population, substitute data, limited features, incorrect output label, biases and injustices in the world, and so on. For example, Bolukbasi et al. (2016) worked on debiasing word embeddings, which are a representation of each word as a vector in about 300 dimensions on the textual context in which the word is found. For example, vectors of words like "queen," "princess," "female," and "woman" would be much closer to each other than other words with them in the embeddings. These word embeddings were trained on Google News articles which would have represented a

woman as one who takes care of the family and a man as the one who earns for the family. The NLP algorithm, however, exhibited gender stereotypes like "A man is to computer programmer as woman is to homemaker." ML algorithms using word embeddings are widespread, which also raises serious concerns of amplifying bias.

Companies and users of ML algorithms can take some precautionary measures so that everyone has an equal opportunity and selection is not based on sensitive attributes. One question that frequently arises is why we don't remove the sensitive attributes and then train the ML algorithms. The problem is that biases can creep in indirectly through other variables like zip code, which can be indicative of individuals' race. In feature engineering (the process of extracting important features from raw data), it is a very important step to remove highly correlated (positive or negative) variables because keeping them would impart the same information to the model and hence would contribute to complexity and possible errors in the model. For example, zip code and percentage of college graduates are highly correlated, and thus one of these variables should be removed.

Apart from descriptive analysis, there are freely available tools and services to evaluate the fairness of ML algorithms. For example, FairTest and Themis enable developers or auditing entities to determine associations between an ML algorithm's sensitive attributes and the benefit/discrimination the algorithm generates. These tools conduct group and individual experiments to determine a link between inputs and outputs. They also discover insights on whether the protective attributes helped in any kind of decision making.

ML algorithms can themselves be used to discover discrimination. Such algorithms include, but are not limited to:

- Classification rule mining: finding association rules between protected attributes and the output with high confidence. For example, if the model, outputs with high confidence a rule that "if a person is from race x, he is not eligible to move to the next round," it will indicate that the decision making was biased and discriminatory toward certain group of people.
- K-nearest neighbor classification: finding similar people with unprotected attributes and checking if protected attributes made the difference in the outcome.
- Bayesian networks: by knowing an event happened, what is the likelihood that one of the known causes was the contributing factor; for example, given the grass is wet, what is the probability it was due to rain or sprinkler?
- Probabilistic causation: increasing the effect of outcome caused due to certain protected attributes everything else being equal.)

One question that frequently arises is why we don't remove the sensitive attributes and then train the ML algorithms. The problem is that biases can creep in indirectly through other variables

Legal Implication Related to the Use of Algorithms

Organizations may face legal liability for the selection tests and processes they use for employee selection, and the use of ML algorithms in hiring poses no exception. In fact, the use of ML algorithms opens a relatively new door in employment law where legal standards that govern decision processes have been outpaced by technology (Kroll et al., 2016).

Specifically, algorithm users should be transparent in the design of the algorithms used in employee selection and should establish a bona-fide auditing process to monitor the performance of algorithms over time.

There are two primary types of discrimination that may arise when ML algorithms are used in the employment context. Disparate treatment discrimination is a form of intentional discrimination. Some forms of disparate treatment discrimination are straightforward in the context of ML algorithms. For instance, disparate treatment discrimination can occur when a protected class characteristic—such as race, sex, religion, national origin, disability or age—is used as one of the input criteria in the algorithm. This is akin to a human decision maker considering an applicant's protected class characteristic when making a selection decision. However, other forms of disparate treatment discrimination are more subtle. Such cases arise when an ML algorithm includes an input that is a proxy for protected class membership, which effectively results in the application of different decision rules to different individuals (Kroll et al., 2016).

Disparate impact discrimination occurs when a neutral employment practice disadvantages one or more groups based on protected class membership. Disparate impact could arise in the context of ML algorithms if one protected class subgroup (e.g., men) consistently outperforms another protected class subgroup (e.g., women) on a given scoring algorithm. Plaintiff applicants demonstrate disparate impact discrimination by presenting statistical evidence to indicate that the resulting performance differences are most likely the result of the scoring algorithm itself and not a result of mere chance or coincidence.

There are certain precautions that organizations can take to reduce their litigation risk when ML algorithms are part of the selection system. Specifically, algorithm users should be transparent in the design of the algorithms used in employee selection and should establish a bona-fide auditing process to monitor the performance of algorithms over time.

Transparency in ML algorithms design involves not only the disclosure of their use in decision making but also transparency in the execution of algorithms. Transparency holds organizations more accountable for the way an ML algorithm performs and the types of inputs it uses to produce outputs. If an organization is transparent about the types of information used to reach decisions and the way an ML algorithm makes decisions, it follows that an organization would specifically focus on ensuring the algorithm is only using job-relevant information to make selection decisions (versus protected class information such as race or sex) and that an algorithm is successfully and consistently identifying candidates most likely to be successful on the job (versus rating candidates at random or failing to identify successful performers at all).

Most notably, the act requires that employers obtain consent from applicants before AI is used to evaluate their interview; AI cannot be used for applicants who do not consent.

Furthermore, recent developments have shifted transparency from something that is recommended to something that is required. Illinois is the first state to regulate the use of ML algorithms in employment interviews (Burstein & DiPrima, 2020). Illinois Artificial Intelligence Video Interview Act ("Video Interview Act"; 2020) requires that applicants are notified that artificial intelligence may be used to analyze the interview and consider their fit for a position. The act also requires that applicants are provided with information before the interview explaining how AI works and in general how it evaluates applicants. Most notably, the act requires that employers obtain consent from applicants before AI is used to evaluate their interview; AI cannot be used for applicants who do not consent.

The General Data Protection Regulation (GDPR), which commenced in May 2018 and provides specific data privacy protections for citizens of the European Union (EU), also contains provisions specific to the use of AI in employee selection. Article 22 of the GDPR prohibits decisions based solely on automated processing that produce legal effects or significantly affects an individual. There are some exceptions to this, most notably applicant consent. Articles 13–15 of the GDPR also state that data subjects must be informed of the existence of automated decision making, though there is controversy over how specific this transparency must be. Regardless, any organization that could potentially have EU citizen applicants must be aware of these transparency requirements as applied to the use of AI in employee selection. Even though GDPR does not apply to non-EU citizens and Illinois is the first U.S. state to formalize transparency in the use of AI in the employment interview process, it is not unlikely that more states will follow suit in the future. Organizations are encouraged to proactively increase the transparency of algorithm design processes regardless of whether they are compelled by law to do so.

Even if organizations take transparent steps to reduce the potential for algorithmic bias in the development stage, it is essential that organizations also audit ML algorithm performance on a regular basis. One of the benefits of AI is that algorithms continuously improve their predictions. However, this also means that an ML algorithm that was not displaying evidence of disparate impact at the outset could begin to have disparate impact as it uses data inputs differently to produce more accurate outputs. As a result, it is imperative that organizations undertake regular reviews of algorithms and make proactive adjustments to avoid the potential for disparate impact liability. How regular these reviews should be is not established by guidelines, but we recommend that they are as frequent as possible and certainly more than once a year. For example, Textio, one of the companies whose algorithms are reviewed in the table above, updates its algorithm constantly to reflect the word patterns of current job postings . This enables Textio to recommend writing a job posting without the unconscious biases as seen "right now" in job postings.

It is not necessary for such oversight to involve the disclosure of source code in its entirety; that is, third parties can audit ML algorithms without concerns of revealing protected intellectual property. For instance, an auditor can examine changes in algorithmic outputs (e.g., an applicant score) when inputs are changed (e.g., stronger or weaker responses to items on an employment test) without full access to the source code. Theoretically, stronger responses on an employment test should yield higher applicant scores, and weaker responses on an employment test should yield lower applicant scores. This is something that can be tested without the need to be privy to everything happening in the code between the input and the output.

Input–output decisions can also be tested in reverse, such as to determine whether the output (e.g. an applicant score) would be the same even if one of the inputs that should not affect the output (e.g., gender, race) were changed (Kroll et al., 2016). This type of oversight based on partial information occurs regularly within the legal system, and ML algorithms—even given their complex nature compared to other selection tools—do not need to be an exception.

Noteworthy, auditing is distinct from canceling or adjusting an applicant's score reactively to avoid disparate impact liability. Such a practice can result in disparate treatment liability, as it involves taking an adverse action against the applicants who took and prepared for the test (e.g., U.S. Supreme Court case *Ricci v. DeStefano,* 2009). Auditing as a proactive strategy for detecting and responding to biased ML algorithms is exactly the type of compliance effort that employment law encourages (Kim, 2017).

The legal recommendations about algorithm transparency, data protection, and auditing do not guarantee algorithmic justice because each case of applying ML algorithms should pass the justice test on its own circumstances, merits, and drawbacks. However, the recommendations are a baseline that HR practitioners should be seeking to establish for a responsible and potentially legally defensible use of ML algorithms in assessment contexts.

Conclusion

In conclusion, we reiterate the basic requirements for algorithmic justice. Providers and users of ML algorithms need to disclose sufficient information about the algorithms so that the algorithms can be independently audited. From both research and workplace law perspectives, a clear and theoretically founded link should be established between the outcome (e.g., predicted job performance) and the algorithmic features, and final assessment scores derived from them.

It is attributed to Albert Einstein to have said that everything should be made as simple as possible but not simpler. Similarly, an ML algorithm should be made as simple as possible for the understanding of its users. Assessed individuals should be able to understand what was measured and, in general terms, how data were modelled by the ML algorithm to arrive at its final recommendations and/or decisions. In sum, from an organizational justice perspective, ML algorithms are just if:

- ML algorithms reliably classify individuals commensurate with their assessed characteristics without using features containing legally protected information.
- Assessed individuals are given the right to know how their characteristics are combined and processed in the ML algorithm. Voicing concerns is allowed too.
- The assessment results are communicated in a personalized and caring way to each individual. AI is still not evolved enough to deliver this feedback and allow a genuine human–human interaction.

References

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1-68.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett, *Advances in neural information processing systems* (Vol. 29, pp. 4349-4357). Barcelona, Spain: Curran.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81*, 77-91.

Burstein, A. & DiPrima, K. (2020). Employers beware: The Illinois Artificial Intelligence Video Interview Act is now in effect. Retrieved from https://www.adlawaccess.com/2020/01/articles/employers-beware-the-illinois-artificial-intelligence-video-interview-act-is-now-in-effect/

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., & Van Gerven, M. (Eds.). (2018). *Explainable and interpretable models in computer vision and machine learning.* Cham, Switzerland: Springer.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the 2nd Conference on Fairness, Accountability and Transparency, AMC*, 329-338.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*(4), 694-734.

Illinois Artificial Intelligence Video Interview Act of 2020, 101-0260.

Kim, P. T. (2017). Auditing algorithms for discrimination. *University of Pennsylvania Law Review, 166,* 189.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *University of Pennsylvania Law Review,* 165, 633.

Liem, C. C. S., Langer, M., Demetriou, A., Hiemstra, A. M. F., Sukma Wicaksana, A., Born, M. Ph., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Cham, Switzerland: Springer.

Next Generation Identification (NGI). (2016, May 6). Retrieved from https://www.fbi.gov/services/cjis/fingerprints-and-other-biometrics/ngi

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia, 16*(4), 1018-1031.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Ricci v. DeStefano, 557 U.S. 557 (2009),129 S. Ct. 2658 (2009).

Society for Industrial and Organizational Psychology. (2019). *Principles for the validation and use of personnel selection procedures* (5th ed.). Bowling Green, OH: SIOP.

Stephen, I. D., Hiew, V., Coetzee, V., Tiddeman, B. P., & Perrett, D. I. (2017). Facial shape analysis identifies valid cues to aspects of physiological health in Caucasian, Asian, and African populations. *Frontiers in Psychology, 8,* 1883.

Volpone, S. D., Tonidandel, S., Avery, D. R., & Castel, S. (2015). Exploring the use of credit scores in selection processes: Beware of adverse impact. *Journal of Business and Psychology, 30*(2), 357-372.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036-1040.