Review Article

Pitfalls of Machine Learning-Based Personnel Selection

Fairness, Transparency, and Data Quality

David Goretzko and Laura Sophia Finja Israel

Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

Abstract. In recent years, machine learning (ML) modeling (often referred to as artificial intelligence) has become increasingly popular for personnel selection purposes. Numerous organizations use ML-based procedures for screening large candidate pools, while some companies try to automate the hiring process as far as possible. Since ML models can handle large sets of predictor variables and are therefore able to incorporate many different data sources (often more than common procedures can consider), they promise a higher predictive accuracy and objectivity in selecting the best candidate than traditional personal selection processes. However, there are some pitfalls and challenges that have to be taken into account when using ML for a sensitive issue as personnel selection. In this paper, we address these major challenges – namely the definition of a valid criterion, transparency regarding collected data and decision mechanisms, algorithmic fairness, changing data conditions, and adequate performance evaluation – and discuss some recommendations for implementing fair, transparent, and accurate ML-based selection algorithms.

Keywords: machine learning, personnel selection, validity, interpretability

Over the past years, more and more organizations have started to use machine learning (ML) modeling (or artificial intelligence; AI[1]) for their recruitment and personnel selection process. Companies like *IBM* and *LinkedIn* use ML algorithms to screen the curriculum vitae (CVs) of their applicants and filter promising candidates automatically (Albert, 2019). Moreover, candidate assessments software like *eSkill* and *testDome* is used (e.g., by *FedEx* and *KPMG*) to determine the suitability of a candidate for a specific position (Kulkarni & Che, 2019). Albert (2019) and Kulkarni and Che (2019) provide overviews of software tools for different steps of the recruiting process and reviews on companies that apply these tools. ML includes various algorithms that can be used to make predictions about new and unseen data after "learning" relationships in fully labeled training data (also known as predictive modeling or supervised learning where a dependent variable is fully observed), to group or cluster observations without labels (unsupervised learning with no observed dependent variable), and to generally detect patterns in complex data (for a gentle introduction to ML modeling, see James et al., 2013). In this paper, we predominantly focus on supervised learning which aims at identifying a function that connects input and output variables (e.g., cognitive ability test scores and the final grades of a trainee) based on observations for which both input and output are known (e.g., human resource data of previous trainees) – the so-called training data. In contrast to classical statistical modeling, ML algorithms are able to handle larger numbers of predictor variables and can often reflect complex interactions and nonlinearities (as they usually make fewer assumptions about the type of relationship). Therefore, ML models often show a high predictive power. For a more detailed discussion of the main differences between ML and classical statistical modeling, see Yarkoni and Westfall (2017).

In the context of recruiting, their ability to process large data sets (especially wide data sets with many [predictor] variables) and to integrate different data sources makes ML models very useful for screening purposes as they can sift through and filter large candidate pools. Amid recent developments such as economy 4.0 and the digitization of

[1] In this paper, we discuss ML applications in personnel selection. These techniques are often referred to as AI for marketing reasons. However, the term AI describes a much broader area of research that tries to develop "intelligent" computer programs and systems without having a clear, stand-alone definition of intelligence (e.g., Wang, 2019). Supervised ML models, as discussed in this paper, only fulfill the definition of a "weak" AI, as these models are only able to solve very narrow tasks. A "strong" or general AI, on the other hand, is defined as an intelligent agent that can flexibly deal with a variety of human tasks (e.g., Etscheid, 2019).

everyday life in general, numerous data sources emerge that can be used to predict various variables of interest – inter alia willingness to buy (Zarazua de Rubens, 2019), employee turnover (Zhao et al., 2019), job satisfaction (Hickman et al., 2019), and employee productivity (Chalfin et al., 2016). Accordingly, a lot of organizations trust in algorithmic procedures to screen applicants (e.g., 72% of the CVs send out to US companies are not examined by humans according to The Guardian; O'Neil, 2016), and many companies have automated large parts of their hiring process (Gonzalez et al., 2019). While organizations already use ML and AI techniques in various steps of recruiting, psychological research still seems to be a bit hesitant when it comes to this topic (König et al., 2020; Putka et al., 2018). However, there are some examples of research with a psychological perspective – for example, Campion et al. (2016) using text-mining and natural language processing techniques for automated candidate screening (automated scoring of candidate essays), Sajjadiani et al. (2019) applying ML algorithms to predict job performance based on work history, Kakulapati et al. (2020) using ML to predict job promotions, or Kang et al. (2020) predicting turnover intention with ML tools.

The automation of the tedious steps of the recruitment and hiring process promises to speed up, simplify, and objectify certain parts of personnel selection, but also comes with some potential downsides (e.g., the creation of discomfort with applicants; Gonzalez et al., 2019) and some severe risks. In 2015, for example, *Amazon* realized that their automated selection algorithm for technical positions favored male applicants by penalizing female resumes (Dastin, 2018). Besides the problem of algorithmic fairness (such as gender bias; Fernández-Martínez & Fernández, 2020), legal restrictions and data protection policies (e.g., Bauer et al., 2020) as well as the dependence of ML modeling on the training data context are challenges that have to be addressed when using ML for automated recruitment.

In this paper, we want to briefly discuss these major challenges and present some ideas on how to tackle them. This selection of major challenges concerning the use of ML modeling for personnel selection does not contain more general issues relevant for proper ML applications (e.g., sample size considerations, hyperparameter tuning, choice of resampling strategies, evaluation metrics, data preprocessing, and model selection – just to mention the most elementary aspects). Table 1 displays these aspects alongside introductory literature that can help to understand how to properly train and evaluate ML models. These methodological considerations are not independent from each other but strongly interrelated and dependent on the application context (e.g., for some model types, feature selection or dimensionality reduction techniques can improve or lower the predictive power; some models

Table 1. Central methodological aspects of ML modeling and introductory literature references

Methodological aspects of ML modeling		
Sample size		e.g., Yarkoni & Westfall (2017)
Model selection	1. Regularized regression	
	2. Tree-based models	
	3. Neural networks	
	4. . . .	e.g., Hastie et al. (2017); James et al. (2013)
Feature selection	1. Preprocessing	
	2. Regularization (e.g., LASSO)	
	3. . . .	
Hyper-parameter tuning	1. Grid search	
	2. Random search	
	3. Bayesian optimization	e.g., Feurer & Hutter (2019)
	4. . . .	
Resampling strategies	1. Cross validation	
	2. Holdout-set	e.g., Hastie et al. (2017)
	3. . . .	
Evalution metrics	1. Classification metrics	e.g., Ferri et al. (2009)
	2. Regression metrics	e.g., Stachl et al. (2020)
. . .		

Note. LASSO = least absolute shrinkage and selection operator; ML = machine learning.

reach a comparably good performance with reasonable hyperparameter defaults, while others strongly call for optimizing the hyperparameters). For further readings on the general challenges of ML that are less specific to personnel selection, we specifically recommend the introductory literature by James et al. (2013) and Yarkoni and Westfall (2017). Hereafter, our main focus is to raise awareness of the pitfalls of ML-based personnel selection (a context that places special demands on ML modeling) as the negative consequences of an invalid selection model can be far-reaching for both companies and applicants.

Challenge 1: Data Quality and Criterion Definition

The first issue with ML-based personnel selection is data quality. Of course, the predictor variables (in ML literature, predictor variables are referred to as features) have to be of a certain quality (e.g., reliable and valid measures of psychological constructs such as personality or cognitive abilities measured with standardized questionnaires and tests) to enable the ML algorithm to learn a substantial relation to the variable of interest (to prevent a *garbage-in-garbage-out problem*). However, since ML algorithms can deal with large sets of predictor variables and have certain measures to prevent overfitting (e.g., regularization or "stopping-rules"), psychological tests and questionnaires as well as questions in job interviews or observations in assessment center exercises do not need to be aggregated to obtain reliable measures (e.g., sum scores or factor scores of a personality questionnaire) but can enter the ML model as separate predictors. Therefore, the quality of single input variables is of less importance compared to traditional selection procedures. Furthermore, new data sources such as video features obtained from automated interviews, automated analyses of assessment center role-plays, presentation on social media, or gamified assessments (e.g., Fernández-Martínez & Fernández, 2020; Woods et al., 2020) can be used on a larger scale (ethical and legal issues should be considered though – see also Challenge 2 and Challenge 3).

The more serious challenge is to define a target variable (or criterion) that the model should predict. In general, three kinds of target variables are possible: (I) a criterion that is based on a definition of the perfect candidate and modeled as a reflective or formative construct from indicators assessed after the candidate was hired (e.g., job performance; MacKenzie et al., 2005), (II) a single key figure such as sales numbers (e.g., Raghavan et al., 2020), or (III) the outcome of the traditional hiring process (a candidate was hired or not – decision rule based on interviews, psychological tests, and/or performance in an assessment center, etc.).

All three options have advantages and downsides. Using a single objective measure (II) like the sales numbers that are usually not influenced by subjective superior's assessments (even though there might be influential factors that are controlled by superiors) may be a good choice to prevent the passing on of prejudices and biases (see also Challenge 3: Fairness). On the other hand, it is difficult to find such an unbiased key figure for each job position as single measures are either overly simplifying (even salespersons should not be judged solely by the volume of their sales) or not to be defined at all (especially for jobs that are not directly related to the company's revenue). In addition, ostensibly objective measures such as sales numbers can also be biased and dependent on the organizational climate (e.g., McKay et al., 2008). If an organization is convinced that the current personnel selection procedure is accurate, it may be meaningful to train an ML model to predict the decision of this procedure whether a candidate fits a position (III). Usually, this approach, which is pure automation of the traditional personnel selection procedure, is not advisable because potential biases in the selection process (and hence in the criterion) might not be apparent (see Challenge 3: Fairness) and further improvements are consequently prevented (the prediction model can be just as good as the criterion it is trained on). Also, retraining such a model will not lead to any improvements, as this will only approximate the initially trained ML model, which itself is an approximation of the original selection procedure.

From a psychological perspective, the first approach (I) is the most natural since it combines ML modeling with a comprehensive definition of the criterion that can be tailored to the requirements of a specific position (for a detailed discussion on how to define such a criterion, see, e.g., Herling, 2000; Van Iddekinge & Ployhart, 2008). Even though this is arguably the most elaborate way to select a target variable, it can be challenging to weigh different aspects of the criterion,[2] and all measures that are deemed to be important for that criterion need to be

[2] Take for example organizational citizenship behavior which can be regarded as a main part of work performance and for which 30 subfacets have been proposed (Campbell & Wiernik, 2015). There are also approaches focusing rather on expertise and competencies than direct performance indicators – concepts that are even more difficult to define (e.g., Herling, 2000).

observed for all subjects in the training data. The latter can be especially difficult as these measures can be obtained only from hired candidates (see also Challenge 5: Evaluation). Moreover, while approaches (II) and (III) are based on rather objective measures and are hence assessed reliably (observable/manifest variables with small measurement error), a newly defined outcome variable as in approach (I) calls for careful construction in line with psychometric guidelines to ensure an objective, reliable, and valid measurement (e.g., Kline, 2015). Accordingly, it should be noted that the less reliable the measurement of the criterion is, the less accurate the predictions of the ML model will be in approximating the true "job performance" (i.e., the true underlying latent variable).

As with a single key figure target variable (II), that is also selected after the hiring process, the selection model can only be trained on the subset of not-rejected candidates. In addition to the loss of data, the criterion can show a rather low variability if the previous selection process was strongly compliant with the criterion. With a low variability of the criterion, the trained model can only reach a low predictive performance. In this case (if the previous selection process is in line with a criterion such as I and II), the simple outcome criterion (III) can be preferred as the model can be trained on the full data set.

Defining a criterion (as in approach I) is not a problem specific to automated personnel selection procedures based on ML, but a common issue of personnel selection in general (e.g., Austin & Villanova, 1992). It is important to highlight that this problem also affects ML-based personnel selection. However, the predictor variables for an ML-based approach do not have to perfectly fit the criterion's definition, as their importance for the prediction of the outcome variable is "learned" by the ML algorithm. In the traditional approaches based on assessment centers and interviews specifically tailored to a prior definition of job performance or expertise, such an intrinsic feature selection is not given and, hence, a valid preselection is even more important.

As stated above, all three presented approaches to define the criterion (or target variable) for the ML modeling have advantages and disadvantages. When the reason for applying ML to the personnel selection process is to simply automate an existing selection process (e.g., to speed up and to save resources in a screening procedure), it seems reasonable to rely on the outcome of the traditional hiring process (III) as the criterion. However, when the aim is to improve the current process, a more nuanced criterion (I) should be used. While this approach can be challenging and limits the evaluation of the model in practice, it seems to be the best way to represent all the important requirements of a job (Van Iddekinge & Ployhart, 2008).

Challenge 2: Transparency

In cases where organizations have decided to use ML modeling and algorithmic procedures to select candidates or at least to screen application documents with these tools, legal issues have to be taken into account. Two aspects are the most prominent – data protection and transparency considering the decision process. The European Union adjusted the legal framework with the General Data Protection Regulation (European Union, 2016), which imposes stricter data storage requirements, demands more open communication on which data are collected, and requests transparent decision making in the context of these algorithmic selection procedures. Liem et al. (2018) describe three parts of *transparency* that have to be fulfilled when rating and selecting applicants. The decision process has to be (1) *comprehensible*, which means that it has to be clear which variables (e.g., test scores, personality traits, or aspects of the CVs) are used for the selection process; (2) *traceable*, which means that it has to be clear how the applicants are rated and therefore why one candidate was ranked above another; and (3) *explainable*, which means that feedback can be given on why a candidate was rejected. All these parts (especially *traceable* and *explainable*) call for a selection model that can be understood by a human. Given the complex architecture[3] of modern ML algorithms (e.g., deep neural networks), it becomes nearly impossible for a human mind to trace back how a trained model comes up with specific predictions. Even simpler models that contain interactions or nonlinear relations are not really *explainable*.

As the problems of explainability and interpretability of ML models have been an increasingly important research topic, several approaches for the so-called interpretable

[3] Modern ML algorithms can reflect rather complex data patterns and relationships due to their increasingly complex architectures. While single elements or parts (e.g., single neurons of a neural network that are rather simple computational units) of the resulting model might still be interpretable by a human observer, the combination of hundreds and thousands of these units (e.g., numerous layers with several neurons in deep neural networks or large numbers of decision trees in tree-based methods such as random forests) prevents the user from understanding how the model comes up with a specific prediction. For this reason, complex ML models are often referred to as black-box models.

ML (IML; Molnar, 2019) or explainable AI (e.g., Hoffman et al., 2018) have been developed. These IML tools may help to build selection models that can live up to the expectations of a *transparent* decision-making procedure. The so-called feature importance or variable importance measures (e.g., Altmann et al., 2010), for example, could be used to make a model *comprehensible* (to some extent) by reflecting which variables were used for the prediction. In combination with a careful variable selection in advance (using only predictor variables that are *comprehensible* themselves), a model that meets the requirement of transparency regarding the variables can be achieved. With global surrogate models (i.e., models that are more easily interpretable are used to approximate the predictions of the full model; Molnar, 2019) and local approximations such as the so-called local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016), it is possible to gain insights into how the model predicts the variable of interest on an algorithmic level.[4] Especially tools like LIME seem to be most promising when it comes to *traceable* and *explainable* personnel selection based on ML. However, the informative value or the interpretability these methods offer is limited by the goodness of the approximation – the better the full model is approximated, the more trustworthy the derived interpretations are. Accordingly, practitioners have to be cautious when applying these IML tools, since complex relations can be reflected incorrectly when the approximating model overly simplifies the black-box model. Even though in most cases more complex ML models reach a higher performance (Fernández Delgado et al., 2014), a simpler model (such as a generalized linear model with regularization) might also yield good performance in some contexts. These more interpretable models naturally show higher *traceability* and *explainability*.

Challenge 3: Algorithmic Fairness

A closely related concern is the fairness of automated personnel selection. The term algorithmic fairness (or fairness in the context of ML) means that the results of an ML algorithm are independent of certain sensitive variables (such as gender, sexuality, or religion) or proxy variables that are strongly related to them (such as zip code that can be related to race in some areas; e.g., Kleinberg et al., 2018; Pessach & Shmueli, 2020). Hence, algorithmic fairness should prevent discrimination and is highly relevant, not only for automating personnel selection procedures but also in contexts like law enforcement, healthcare, or education. Pessach and Shmueli (2020) provide an overview of various measures of fairness in ML. Often these measures in some circumstances cannot be optimized simultaneously.

If the selection procedure is not accessible and transparent, it is impossible to ensure fairness (e.g., gender, race, age, or religion). However, even if we can provide methods that allow for a transparent automated personnel selection, it remains difficult to guarantee a fair procedure. As stated above (Challenge 1: Data Quality) depending on the data quality and the training data conditions more generally, it is possible that prejudices and biases are inherited from the nonautomated personnel selection procedure. For example, if a superior's assessments in an organization are racially biased and an ML model "learns" that good applicants are those that most likely will get a positive review by a superior, then the model will also "learn" to discriminate against candidates based on race. Both predictor and target variables can contain bias – although it is more severe when the target variable is biased, as an ML model cannot adopt this bias if the predictors are completely unrelated to discriminatory variables such as gender or race.

Besides "inheriting" biases, imbalanced training data can also yield bias in the prediction model. For example, an IT company might mainly have male employees, and consequently, the training data will be strongly unbalanced with fewer data from female applicants. An ML model trained on this unbalanced data set will potentially classify female applicants less accurately or give more weight to more typical male traits. There are several more examples of how the training data context can induce bias (e.g., Lee, 2018). The issue of predictive bias or differential prediction (and related under- or overprediction of job performance) is well-known in personnel selection research (e.g., Berry, 2015). It describes cases where a grouping variable (e.g., gender) moderates the effect of a predictor variable (usually cognitive ability test scores) on the job-related outcome variable (e.g., Berry, 2015). These biases can be

[4] LIME is a method to explain how the complex ML model comes up with an individual prediction. In the context of ML-based personnel selection, it could be used to find out why a candidate (i.e., a single observation in the data set) is deemed suitable for a position or not by the trained ML model. The idea of LIME is to approximate a complex decision model (ML model) with a linear (and therefore interpretable) function locally (which means in close proximity to the observation of interest). To this end, several artificial observations are created that are similar to the observation that should be explained. The outcome variable of these artificial observations is predicted based on the complex ML model and a linear (or logistic regression) is fitted to these sampled observations and their predicted outcome. The resulting interpretable model can then be used to explain the "decision-making" of the ML model for the observation of interest.

relatively easily diagnosed when the relations between few predictor variables and the criterion are assessed with regression analyses (Berry, 2015; Meade & Fetzer, 2009; Niessen et al., 2019). When using ML modeling, on the other hand, the often large amount of predictor variables, complex interactions, and the highly nonlinear links to the criterion can make it incomparably more difficult to detect differential predictions.

Even (small, but) representative samples can cause problems when the ML model is trained by optimizing the overall performance (using default cost functions) since it may neglect minorities. One potential way to handle this problem is oversampling – a strategy that is usually applied to data sets with unbalanced target variables (i.e., data where the categorical dependent variable is unevenly distributed) where sampling strategies are used to enlarge smaller classes (e.g., Sáez et al., 2016).

As Zou and Schiebinger (2018) point out, there are two general ways to deal with the discussed issues of fairness. First, algorithms have to be developed that are robust against imbalanced training data, for example, by adding constraints that implement a parity principle which can be understood as quota rules (with regard to gender, race, or age) and variable selection procedures that prevent gender, race, etc. related variables from entering the model (e.g., Grgic-Hlaca et al., 2016). Regarding data quality, ML and IML methods can be used to detect biases in current selection procedures and hence in the training data. Zou and Schiebinger (2018) call an ML model that identifies stereotypes and biases *AI auditor*, but you could also refer to this approach as *ML-based reverse engineering*. If you use ML to identify issues like gender bias by predicting gender based on the variables of the selection model and this reverse engineering ML model shows no relation between the predictor variables and gender, the selection model based on the same set of variables will not be biased with regard to gender (or analogously race, etc.).

Although there are tools to detect and mitigate biases that affect minorities, it remains challenging to create fair and accurate automated personnel selection models because often variables like race/ethnicity, religion, and also sometimes gender are not collected. Even though this is often done with the intent to achieve fairness in the selection process, it makes it hard to determine if these categories affected the selection nonetheless. Furthermore, practitioners have to be aware that current practices that directly impact the target variable of the ML model can be biased. In psychological research, it was shown that the stereotype threat can have adverse effects on test performance and therefore on the personnel selection process (e.g., Casad & Bryant, 2016) and that gender stereotypes are still prominent and influential in this context (e.g., Heilman et al., 2015). Hence, to prevent these biases and to ensure fairness in automated personnel selection, we have to question the training data, the algorithms, and the human factor as well. As presented in Challenge 2, IML tools are helpful to understand how the trained ML model comes up with its predictions. For example, when combining these tools with data-driven personas (e.g., Salminen et al., 2020) which can be described as aggregated representations (e.g., over a number of variables) that characterize a group of people, one can create a model-agnostic method to evaluate the algorithmic fairness of models that are already applied for personnel selection.

Challenge 4: Changing Data Conditions

As an automated ML selection model is dependent on the data quality and the specific data conditions it is trained on (see Challenge 1: Data Quality), two problems arise – generalizability and robustness against changing data conditions. The generalizability problem which is a challenge for all predictive models can be seen as the smaller problem as ML personnel selection models are usually trained on data of one organization and are therefore tailored to the specific data context they are then applied to. Moreover, the performance of an ML model (or any other predictive model) should always be evaluated on out-of-sample data (i.e., data that the model was not trained on). As new data cannot always be collected easily, resampling and cross-validation strategies are applied to approximate the model performance under new data conditions (for an introduction on this topic, see James et al., 2013; Putka et al., 2018). Furthermore, it might be useful to evaluate the applicability of an ML model to new empirical data, so that one knows whether the model is trustworthy in a specific context. For this reason, so-called applicability domain analyses (ADA) have emerged. ADA for ML is quite popular in the context of molecular chemistry (Sahigara et al., 2012) but becomes more and more common in other fields as well (lately, an R package has been made available that covers some applicability measures; Gotti & Kuhn, 2020). With ADA, one can determine how similar a new data set (e.g., data of new applicants in a company) is to the training data (e.g., data of previously hired employees) and, hence, if the predictions of the trained model can be trusted.

Changing data conditions and the so-called concept drift are a more severe problem. Concept drift describes a shift of meaning in a variable (or a combination of variables)

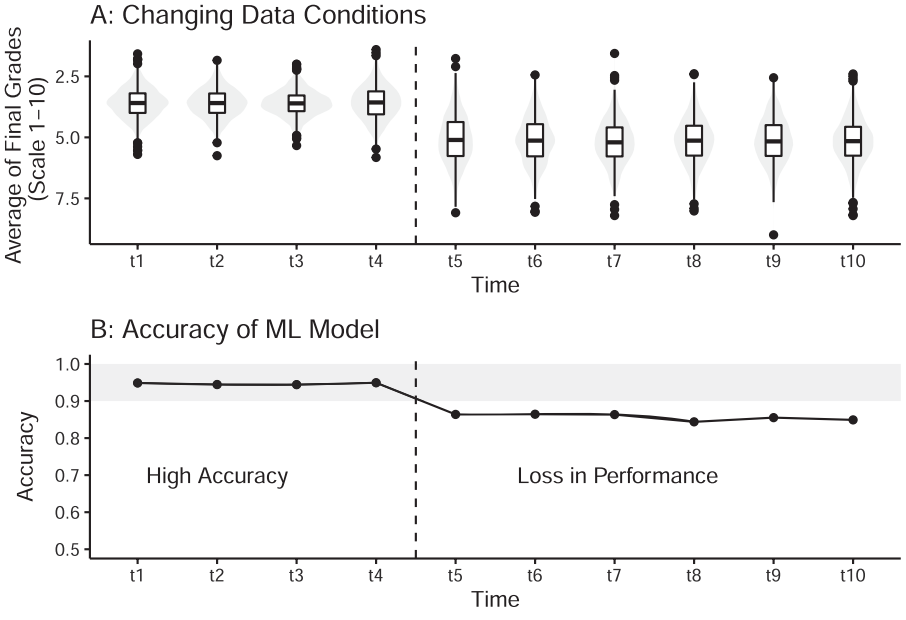

Figure 1. Exemplary visualization: Changing data conditions and concept drift (A) and related drop in performance (B) of the ML selection model (considering an accuracy of at least 90% as acceptable).

that can deteriorate the predictive performance of a model. The issue can be illustrated by the following scenario that is presented in Figure 1: Let us assume that a prediction model is used to classify good (or suitable) and poor candidates for a specific trainee program and that the most important predictors of the model are the final grades of the candidates. Due to homeschooling programs and short-term closures of schools during the COVID-19 pandemic (after t4), however, final grades might decrease on average. Moreover, students from families with a lower socioeconomic status might be more strongly affected by this trend (e.g., due to lack of technological equipment). Consequently, the relation between the final grades and suitability for the trainee program will be distorted since school grades no longer reflect the students' skills and knowledge to the same extent. One can easily imagine that in this scenario the performance of a model that is trained on the first cohort (t1; pre-COVID-19) will decrease when applied to predict the suitability of later cohorts as the meaning of the final grades changed. Figure 1B illustrates how the accuracy of the selection ML falls below 90% (the desired lower bound in this example), as the average grades drop drastically after t4. The grades are no longer a valid predictor of job performance as they become a less valid measure of skills and knowledge.

As this example shows, changing data conditions or concept drift can be a severe problem when applying a stationary ML model to select candidates. When dealing with this issue, the first step is to detect a problem – either by identifying changes in important predictor variables or by observing a loss in performance (the evaluations of such models can be challenging though; see Challenge 5: Evaluation). Either way, automated personnel selection models have to be monitored constantly, so that the people in charge can intervene immediately. There are approaches (for an overview, see Hoens et al., 2012) that integrate the detection of change or concept drift in the ML model itself by weighing the different models in an ensemble according to their ability to accurately predict difficult (or changing) instances (e.g., Street & Kim, 2001) or by weighing the training data or rather using only parts of the data with windowing techniques (e.g., Lazarescu et al., 2004). As these algorithms often need larger samples, it can be necessary for practitioners to rely on manual adjustments, which means monitoring the data as well as the model performance and retraining the model in case a drift is detected. If the model training is not too computationally costly, updating the training set and retraining the model regularly is advisable. When retraining the model, it seems promising to put higher weights on more recent data to mitigate the negative effects of a changing data context.

Challenge 5: Evaluation

Another problem emerges when it comes to constantly monitoring and evaluating the performance of an ML model in this specific application context. If the selection process is completely automated and the decision of which candidate is hired (and which candidate is not) is solely based on the predicted outcome of an ML model, the evaluation of the selection model will become challenging

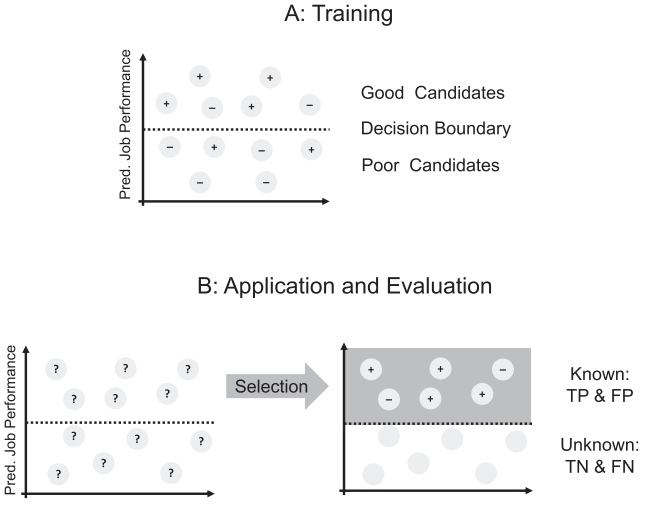

Figure 2. Exemplary visualization: training of ML model (A) as well as application to new data and limits of evaluation (B; TP, TN, FP and FN stand for TP/FP/FN). FN = false negatives; FP = false positives; ML = machine learning, TN = true negatives; TP = true positives.

as the possible performance measures will be limited. Since rejected candidates are no longer observable, *true negatives* (TN) and *false negatives* (FN) are not available for evaluation. This can be seen as an extreme form of variance restriction in the criterion, which is a familiar problem in psychological research in general (e.g., Sackett & Yang, 2000; Westreich, 2012) and in the personnel selection research in particular (Schmidt et al., 2006).

Figure 2 illustrates this issue. The training data consist of candidates that were selected based on a traditional selection procedure for which the true outcome (i.e., the true job performance measure) was observed. The ML model is trained to optimally differentiate between good and poor candidates (Figure 2A). The trained personnel selection model and its decision boundary are then applied to a new pool of candidates (Figure 2B). Based on the model's predictions, only candidates classified as suitable will be hired. For the selected candidates, we can assess their performance on the job. For the rest of the candidate pool, however, we do not have this information.

Therefore, if the selection of candidates is based on the ML model without exceptions, one has to rely on the positive predictive value (also known as the precision), which is defined as the proportion of *true positives* (TP) in all positives (TP + *false positives*, FP). Other performance measures that would be of interest, such as the overall accuracy, the specificity and the sensitivity of the selection procedure cannot be evaluated (due to the missing information on TN and FN). In cases where, for example, the costs of hiring are extremely high and FP have to be avoided, one would focus on the *false positive rate* ($FPR = \frac{FP}{FP+TN}$), which cannot be calculated either. Hence, practitioners have to be aware that the possibilities to evaluate a trained ML model (or any other decision rule) that is used to select the candidates in the first place (and therefore determines the evaluation data) are rather limited. To ensure a proper evaluation, a little percentage of candidates that were rejected by the ML model would have to be hired. However, this seems to be a rather uneconomical and uncertain solution for employers. Accordingly, the problem of evaluating a selection process based on observations filtered by this exact selection process remains a critical issue for personnel selection practice. For this reason, the knowledge and experience of industrial and organizational (I-O) psychologists are necessary to question the choices of the selection model/decision rule (regardless of whether it is a traditional approach or an ML model) and randomly check the decisions for or against a candidate for content validity. In doing so, IML tools (see also Challenge 2) might help to understand how a decision was made.

Summary

In addition to general issues related to ML modeling (e.g., proper evaluation, variable selection, or parameter tuning), the five challenges presented above are factors that have to be taken into account when using ML tools for personnel selection processes. With this paper, we want to raise awareness that ML modeling does not magically solve common problems of personnel selection (e.g., the evaluation of a selection process based on data that were selected by that process – see Challenge 5) and still needs human experts to monitor it (see Challenge 4).

However, as it promises to make the recruitment processes more efficient, which can lead to clear financial benefits (e.g., Ben-Gal, 2019), ML-based personnel selection and human resource (HR) management will become even more popular (Kulkarni & Che, 2019). Therefore, practitioners should take the presented challenges seriously when applying ML in their recruitment and personnel selection processes.

When using ML models in personnel selection, it is important to ensure that:

1. The data used to train the model is selected carefully – especially that a criterion is defined that reflects all job requirements (Van Iddekinge & Ployhart, 2008).
2. The decisions based on the ML model are transparent (i.e., *comprehensible*, *traceable*, and *explainable*). If so-called "black-box" models are used due to higher accuracy, model-agnostic IML tools (Molnar, 2019) like surrogate models or LIME may help to achieve this goal.

3. Algorithmic fairness is given. IML tools in combination with specifically designed personas as well as *AI auditors* (models that try to identify biases; see Zou & Schiebinger, 2018) can be used to evaluate algorithmic fairness in this context.
4. The model is monitored regularly to detect changing data conditions and issues such as *concept drift*. Weighting training data (and favoring newer data) can soften the negative effects of concept drift.
5. Human expertise (I-O psychologists and HR experts) is used to evaluate the model's decisions as selection bias (or variance restrictions) can limit the applicability of common performance measures (e.g., sensitivity and specificity). Again, IML tools may help to understand and evaluate the model predictions.

Conclusion

ML models and their high predictive power are a promising tool when it comes to the automation of personnel selection or parts of the recruitment process. When used correctly, the selection of good candidates can be achieved in a much more economical way for employers (compared to classical recruitment procedures) by screening a large candidate pool efficiently and by predicting objective job performance scores with high accuracy. In theory, automated personnel selection can be applied to every type of position, given that meaningful predictor variables and valid outcome variables can be determined – and as long as enough data for this specific job profile are available to train the model. In practice, the applicability of the approach might vary among jobs (e.g., jobs that require stronger social skills which are harder to quantify compared to jobs that rely on specific domain knowledge that can be reliably assessed with standardized tests) and levels of expertise (e.g., high profile jobs might be more unique in terms of their job requirements and, hence, enough data to train an ML model might be available less often). Hence, ML modeling and automated personnel selection seem to be especially useful in contexts where several similar positions have to be filled regularly (e.g., trainee programs) and less appealing when it comes to hiring specialists for very unique positions. Accordingly, organizations can benefit from ML modeling when they have enough relevant data, that is, they have a large candidate pool to choose from and numerous vacancies with similar job requirements. For this reason, mainly rather large companies adopt ML-based recruitment strategies (Albert, 2019).

For candidates, on the other hand, ML-based selection models have the potential to make the selection process fairer by reducing biases that might be apparent in a more traditional personnel selection process. As we tried to emphasize with the present paper though, these possible benefits of ML models are not intrinsic to the method itself. Instead, practitioners (i.e., HR officers) face several challenges that can harm the validity and quality of the selection procedure. To achieve a trustworthy ML selection algorithm, the data basis and the model have to be carefully inspected and evaluated. One has to be cautious when defining the selection criterion, transparency and fairness have to be preserved and the ML selection models should be monitored regularly to detect changing data conditions that can influence the model's performance. When ML is used inappropriately, it can severely hurt organizations and candidates, so overly optimistic promises made by providers of respective software solutions should always be questioned. Hence, practitioners need a fundamental understanding of ML modeling as well as domain-specific knowledge to develop fair, transparent, and accurate selection procedures. With the present paper, we hope to make the possible pitfalls of automated personnel selection more apparent and ask practitioners for a careful and thoughtful application of ML models as automated selection procedures can have far-reaching consequences for employers and candidates.

References

Albert, E. T. (2019). AI in talent acquisition: A review of AI-applications used in recruitment and selection. *Strategic HR Review*, *18*(5), 215–221. https://doi.org/10.1108/SHR-04-2019-0024

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology*, *77*(6), 836–874. https://doi.org/10.1037/0021-9010.77.6.836

Bauer, T. N., Truxillo, D. M., Jones, M. P., & Brady, G. (2020). Privacy and cybersecurity challenges, opportunities, and recommendations: Personnel selection in an era of online application systems and big data. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychology* (pp. 393–409). American Psychological Association. https://doi.org/10.1037/0000193-018

Ben-Gal, H. C. (2019). An ROI-based review of HR analytics: Practical implementation tools. *Personnel Review*, *48*(6), 1429–1448. https://doi.org/10.1108/PR-11-2017-0362

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 47–74. https://doi.org/10.1146/annurev-orgpsych-032414-111427

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate

essays for personnel selection. *Journal of Applied Psychology*, *101*(7), 958–975. https://doi.org/10.1037/apl0000108

Casad, B. J., & Bryant, W. J. (2016). Addressing stereotype threat is critical to diversity and inclusion in organizational psychology. *Frontiers in Psychology*, *7*, 8. https://doi.org/10.3389/fpsyg.2016.00008

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124–127. https://doi.org/10.1257/aer.p20161029

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

Etscheid, J. (2019). Artificial intelligence in public administration. In I. Lindgren, M. Janssen, H. Lee, A. Polini, M. P. Rodríguez Bolívar, H. J. Scholl, & E. Tambouris (Eds.), *Electronic government* (pp. 248–261). Springer International Publishing. https://doi.org/10.1007/978-3-030-27325-5_g19

Council of the European Union. (2016, April). General data protection regulation. https://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf

Fernández Delgado, M., Cernadas Garcia, E., Barro Ameneiro, S., & Amorim, D. G. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*(1), 3133–3181. https://doi.org/10.5555/2627435.2697065

Fernández-Martínez, C., & Fernández, A. (2020). AI and recruiting software: Ethical and legal implications. *Paladyn, Journal of Behavioral Robotics*, *11*(1), 199–216. https://doi.org/10.1515/pjbr-2020-0030

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated machine learning* (pp. 3–33). Springer.

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). "Where's the IO?" Artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions*, *5*(3), 4–12. https://doi.org/10.25035/pad.2019.03.005

Gotti, M., & Kuhn, M. (2020). *Applicable: A compilation of applicability domain methods*. https://CRAN.R-project.org/package=applicable

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016, December 3–8). *The case for process fairness in learning: Feature selection for fair decision making* [Paper presentation]. Symposium on machine learning and the law at the 29th Conference on Neural Information Processing Systems, Barcelona, Spain. https://doi.org/https://www.mlandthelaw.org/papers/grgic.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning* (Vol. 2). Springer series in Statistics.

Heilman, M. E., Manzi, F., & Braun, S. (2015). Presumed incompetent: Perceived lack of fit and gender bias in recruitment and selection. In A. M. Broadbridge & S. L. Fielden (Eds.), *Handbook of gendered careers in management* (pp. 90–104). Edward Elgar Publishing. https://doi.org/10.4337/9781782547709.00014

Herling, R. W. (2000). Operational definitions of expertise and competence. *Advances in Developing Human Resources*, *2*(1), 8–21. https://doi.org/10.1177/152342230000200103

Hickman, L., Saha, K., De Choudhury, M., & Tay, L. (2019). Automated tracking of components of job satisfaction via text mining of twitter data. In *ML Symposium, SIOP*.

Hoens, T. R., Polikar, R., & Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: An overview. *Progress in Artificial Intelligence*, *1*(1), 89–101. https://doi.org/10.1007/s13748-011-0008-0

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining explanation for "Explainable AI". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 197–201. https://doi.org/10.1177/1541931218621047

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Kakulapati, V., Chaitanya, K. K., Chaitanya, K. V. G., & Akshay, P. (2020). Predictive analytics of HR – A machine learning approach. *Journal of Statistics and Management Systems*, *23*(6), 959–969. https://doi.org/10.1080/09720510.2020.1799497

Kang, I. G., Croft, B., & Bichelmeyer, B. A. (2020). Predictors of turnover intention in U.S. Federal Government Workforce: Machine learning evidence that perceived comprehensive HR practices predict turnover intention. *Public Personnel Management*. Advance online publication. https://doi.org/10.1177/0091026020977562

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings*, *108*, 22–27. https://doi.org/10.1257/pandp.20181018

Kline, P. (2015). *A handbook of test construction (psychology revivals): Introduction to psychometric design*. Routledge.

König, C. J., Demetriou, A. M., Glock, P., Hiemstra, A. M., Iliescu, D., Ionescu, C., & others (2020). Some advice for psychologists who want to work with computer scientists on big data. *Personnel Assessment and Decisions*, *6*(1), 2. https://doi.org/10.25035/pad.2020.01.002

Kulkarni, S. B., & Che, X. (2019). Intelligent software tools for recruiting. *Journal of International Technology and Information Management*, *28*(2), 2–16. https://scholarworks.lib.csusb.edu/jitim/vol28/iss2/1

Lazarescu, M. M., Venkatesh, S., & Bui, H. H. (2004). Using multiple windows to track concept drift. *Intelligent Data Analysis*, *8*(1), 29–59. https://doi.org/10.3233/IDA-2004-8103

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, *16*(3), 252–260. https://doi.org/10.1108/JICES-06-2018-0056

Liem, C. C. S., Langer, M., Demetriou, A., Hiemstra, A. M. F., Sukma Wicaksana, A., Born, M. P., & 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). Springer International Publishing. https://doi.org/10.1007/978-3-319-98131-4_g9

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, *90*(4), 710–730. https://doi.org/10.1037/0021-9010.90.4.710

McKay, P. F., Avery, D. R., & Morris, M. A. (2008). Mean racial-ethnic differences in employee sales performance: The moderating role of diversity climate. *Personnel Psychology*, *61*(2), 349–374. https://doi.org/10.1111/j.1744-6570.2008.00116.x

Meade, A. W., & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods*, *12*(4), 738–761. https://doi.org/10.1177/1094428109331487

Molnar, C. (2019). *Interpretable machine learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2019). Gender-based differential prediction by curriculum samples for college admissions. *Educational Measurement: Issues and Practice*, *38*(3), 33–45. https://doi.org/10.1111/emip.12266

O'Neil, C. (2016). *How algorithms rule our working lives*. The Guardian. https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives

Pessach, D., & Shmueli, E. (2020). *Algorithmic fairness*. arXiv. https://arxiv.org/pdf/2001.09784.pdf.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689–732. https://doi.org/10.1177/1094428117697041

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 469–481).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, *57*, 164–178. https://doi.org/h10.1016/j.patcog.2016.03.012

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, *17*(5), 4791–4810. https://doi.org/10.3390/molecules17054791

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225. https://doi.org/10.1037/apl0000405

Salminen, J., Jung, S.-g., Chowdhury, S. A., & Jansen, B. J. (2020). Rethinking personas for fairness: Algorithmic transparency and accountability in data-driven personas. In H. Degen & L. Reinerman-Jones (Eds.), *Artificial intelligence in HCI* (pp. 82–100). Springer International Publishing.

Schmidt, F. L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology*, *59*(2), 281–305. https://doi.org/10.1111/j.1744-6570.2006.00065.x

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*, *34*(5), 613–631. https://doi.org/10.1002/per.2257

Street, W. N., & Kim, Y. (2001). A streaming ensemble algorithm (SEA) for large-scale classification. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 377–382). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/502512.502568

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology*, *61*(4), 871–925. https://doi.org/10.1111/j.1744-6570.2008.00133.x

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology*, *23*(1), 159–164. https://doi.org/10.1097/EDE.0b013e31823b6296

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, *29*(1), 64–77. https://doi.org/10.1080/1359432X.2019.1681401

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zarazua de Rubens, G. (2019). Who will buy electric vehicles after early adopters? Using machine learning to identify the electric vehicle mainstream market. *Energy*, *172*, 243–254. https://doi.org/10.1016/j.energy.2019.01.114

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2019). Employee turnover prediction with machine learning: A reliable approach. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Intelligent systems and applications* (pp. 737–758). Springer International Publishing. https://doi.org/10.1007/978-3-030-01057-7_g56

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist – It's time to make it fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

History
Received October 12, 2020
Revision received May 29, 2021
Accepted June 20, 2021
Published online October 20, 2021

David Goretzko
Department of Psychology
Ludwig-Maximilians-Universität München
Leopoldstr. 13
80802 Munich
Germany
david.goretzko@psy.lmu.de