https://doi.org/10.1007/s12599-020-00673-w

RESEARCH PAPER

Highly Accurate, But Still Discriminatory

A Fairness Evaluation of Algorithmic Video Analysis in the Recruitment Context

Alina Köchling · Shirin Riazy · Marius Claus Wehner · Katharina Simbeck

Received: 28 February 2020 / Accepted: 19 October 2020

Abstract The study aims to identify whether algorithmic decision making leads to unfair (i.e., unequal) treatment of certain protected groups in the recruitment context. Firms increasingly implement algorithmic decision making to save costs and increase efficiency. Moreover, algorithmic decision making is considered to be fairer than human decisions due to social prejudices. Recent publications, however, imply that the fairness of algorithmic decision making is not necessarily given. Therefore, to investigate this further, highly accurate algorithms were used to analyze a pre-existing data set of 10,000 video clips of individuals in self-presentation settings. The analysis shows that the under-representation concerning gender and ethnicity in the training data set leads to an unpredictable overestimation and/or underestimation of the likelihood of inviting representatives of these groups to a job interview. Furthermore, algorithms replicate the existing inequalities in the data set. Firms have to be careful when implementing algorithmic video analysis during recruitment as biases occur if the underlying training data set is unbalanced.

Keywords Fairness · Bias · Artificial algorithm decision making · Recruitment · Asynchronous video interview · Ethics · HR analytics · Artificial intelligence

Accepted after two revisions by the editors of the special issue.

A. Köchling (✉) · M. C. Wehner
Heinrich-Heine-University, Düsseldorf, Germany
e-mail: alina.koechling@hhu.de

S. Riazy · K. Simbeck
HTW Berlin, Berlin, Germany

1 Introduction

Currently, among recruitment functions, a global wave of enthusiasm is arising about algorithmic decision making in the context of recruitment and job interviews (Langer et al. 2019; Persson 2016). Here, algorithmic decision making can be understood as automated decision making and remote control as well as standardization of routinized decisions in the workplace (Möhlmann and Zalmanson 2017). One often-used application of HR analytics in the recruiting context is algorithmic video analysis, where firms receive an evaluation of each applicant and a prediction of the applicants' job performance. The algorithmic video analysis takes place asynchronously; the applicants record a video of themselves, which is then algorithmically evaluated (Langer et al. 2019; Dahm and Dregger 2019). Limited time and resources of recruiters simultaneously managing large pools of applicants are some of the main reasons for the rapid growth of algorithmic decision making in many companies (Leicht-Deobald et al. 2019). Algorithmic decision making in recruitment is presently well-established in large companies from a variety of industries, such as Vodafone, KPMG, BASF, and Unilever (Daugherty and Wilson 2018). It has both practical and economic benefits as recruiters become more efficient in handling and screening applicants in less time, which, in turn, reduces the time-to-hire and increases the speed of the entire recruitment process (Suen et al. 2019).

Moreover, firms want to increase the objectivity and fairness of the recruitment process by implementing algorithmic decision making and seeking to diminish human bias (e.g., prejudices and personal beliefs) (He 2018). In computer science, two types of fairness can be distinguished: group fairness and individual fairness (Zemel et al. 2013). Group fairness, which is also known as

Published online: 24 November 2020

statistical parity, ensures that overall positive (negative) classifications are similar for protected groups and the overall population (see, e.g., Kamishima et al. (2012)). Individual fairness ensures that any two individuals who are "similar" should be classified similarly (Dwork et al. 2012; Zehlike et al. 2020). Concerning group fairness, operationalizations of fairness measurements are closely connected and often equal to inter-group differences in measures of accuracy (Friedler et al. 2019). These types of algorithms have been considered to be biased because human biases were transferred to the algorithm (Barocas and Selbst 2016), thereby making them "unfair" (Mehrabi et al. 2019). This definition is closely related to the statistical bias, which is defined as the systematic error (or tendency) of an estimator (Kauermann and Kuechenhoff 2010).

There are several factors which may lead to biased algorithms and, in turn, unfairness. A natural cause of a biased algorithm is biased input data, which may contain explicit or implicit human judgments and stereotypes (Diakopoulos 2015; Suresh and Guttag 2019). Bias may also occur if the data are inaccurate (Kim 2016) or if there is a "mismatch between users and system design" (see Table 1 in Friedman and Nissenbaum (1996), p. 335). Moreover, Shankar et al. (2017) discussed the representation bias in the ImageNet and Open Images data sets, where the representation imbalance led to a decreased relative performance.

Table 1 Means and standard deviations of personality trait values in the First Impressions V2 data set and the classifications of the test/validation set (n = 2000)

Trait	Asian	Caucasian	African-American
Training data set			
Job interview score	0.52 ± 0.13	0.51 ± 0.15	0.48 ± 0.14
Conscientiousness	0.54 ± 0.14	0.53 ± 0.16	0.49 ± 0.15
Neuroticism	0.47 ± 0.13	0.48 ± 0.15	0.50 ± 0.15
Test data set			
Job interview score	0.54 ± 0.10	0.51 ± 0.15	0.46 ± 0.13
Conscientiousness	0.54 ± 0.10	0.53 ± 0.15	0.49 ± 0.14
Neuroticism	0.45 ± 0.11	0.47 ± 0.16	0.52 ± 0.14
BU-NKU (test set)			
Job interview score	0.50 ± 0.08	0.51 ± 0.11	0.47 ± 0.10
Conscientiousness	0.54 ± 0.10	0.53 ± 0.15	0.49 ± 0.14
Neuroticism	0.45 ± 0.11	0.47 ± 0.16	0.52 ± 0.14
ROCHCI (test set)			
Job interview score	0.50 ± 0.07	0.51 ± 0.07	0.49 ± 0.06
Conscientiousness	0.54 ± 0.10	0.53 ± 0.15	0.49 ± 0.14
Neuroticism	0.45 ± 0.11	0.47 ± 0.16	0.52 ± 0.14

It is well-known that impression plays a vital role during the selection process because recruiters make their conclusions based on their impression of the candidate's personality and the person-organization fit (Barrick et al. 2010). However, interviewers tend to base their decisions on limited information from those impressions (Anderson 1960; Springbett 1958; Frieder et al. 2016), known as subjective human bias. Several subjective aspects might influence the perception during the interview, such as applicants' appearance, ethnicity, gender, or age (Lepri et al. 2018; Levashina et al. 2014; Schmid Mast et al. 2011). Hence, in addition to cost reduction and efficiency reasons, companies want to avoid an implicit subjective human bias by using algorithmic decision making to increase the objectivity and fairness of the recruitment process (Langer et al. 2019; Persson 2016).

Several providers offer algorithmic selection tools, such as the American company HireVue and the German company Precire. In Germany, more than 100 companies used Precire´s algorithmic assistance in 2018 (Precire 2020). While these service providers offer support in handling and screening applications more efficiently, they are also claiming to provide psychological profiles of the candidates, such as personality traits (e.g., conscientiousness and emotional stability) which are associated with job performance (Barrick et al. 2010; Linnenbürger et al. 2018).

Despite the enthusiasm for algorithmic decision making in the recruiting context, there remain concerns regarding the possible threat of unfairness by relying solely on algorithmic decision making (Lee 2018; Lindebaum et al. 2019). The unfair implicit treatment could, for example, jeopardize diversity among employees, which has increasingly become a business priority (Economist 2019). Algorithms are often highly accurate, and "the accuracy performance of apparent personality recognition models is generally measured in terms of how close the outcomes of the approach to the judgments made by external observers (i.e., annotators) are" (Junior et al. 2019, p. 3). However, with preferential sampling and implicit biases of the training data, discriminatory tendencies could be replicated, systematically discriminating against a subgroup (Calders and Verwer 2010; Calders and Žliobaitė 2013). This concern leads to our particular research question: Despite a high accuracy, what changes occur to the likelihood to be invited for a job interview when there is an unequal distribution of groups? While it is clear that biased data lead to biased results, this paper aims to discuss the imbalanced representation problem in the context of fairness, which is still a research gap. Specifically, we go beyond the Shankar et al. (2017) findings by analyzing the nature and relevance of deviations in the classifications in an HR context.

Therefore, the aim of this study is fourfold. First, we examine whether algorithms reinforce existing inequalities in their training data sets, specifically in the recruiting context. Second, we examine whether an underrepresentation of certain groups (e.g., gender, ethnicity) leads to unpredictable classifications for those groups when there was no unfairness previously (see Sect. 2.3.3, where the representation imbalance is introduced). Our results carry important implications for the hiring process because the findings raise doubts about the objectivity and fairness of algorithmic decision making if the training data set contains inequalities or unknown biases. Third, we contribute to the current debate on ethical issues associated with HR analytics and algorithmic decision making, including bias and unfairness (Barocas and Selbst 2016; Lepri et al. 2018), since there are only a few published academic articles and knowledge on the potential pitfalls of HR analytics is still limited (Marler and Boudreau 2017; Mehrabi et al. 2019).

Furthermore, we contribute to the computer science literature by providing a representative example in which the algorithms reinforce existing biases and where an underrepresentation leads to unpredictable classifications. This is to be handled separately from the class imbalance problem, as discussed in Sect. 2.3.3.

To answer our question, we applied an exploratory approach and used an existing data set of the *ChaLearn Looking at People First Impression V2 challenge* consisting of 10,000 15-s video clips and two winning algorithms. Since a classical hypothesis test would not be appropriate at this point, we conducted a criterion evaluation using methods from computer science (Dwork et al. 2012; Hardt et al. 2016; Feldman et al. 2015). The videos included the five-factor model (FFM) personality traits and an indication of whether the person should be invited (Ponce-López et al. 2016). As conscientiousness and neuroticism are influential predictors for job performance (Barrick et al. 2001), our study focuses on these two personality traits of the FFM. With the application of machine learning to people-related data, we contribute to the evaluation and validation of video analysis in recruitment.

2 Theoretical Background

2.1 Personality Trait Inference

Since stable individual characteristics are indicators for behavioral patterns, personality is a valid predictor of job performance, among other criteria (Hurtz and Donovan 2000; Vinciarelli and Mohammadi 2014). Previous research has shown that personality traits are influential factors for employment interview outcomes (Huffcutt et al. 2001). The FFM is the dominant personality framework for personnel selection and consists of five dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism (Rothstein and Goffin 2006).

Especially conscientiousness and neuroticism are considered valid predictors of job performance (Barrick and Mount 1991; Barrick et al. 2010; Hurtz and Donovan 2000; Behling 1998). Conscientiousness is, across all situations and activities, the strongest predictor of general job performance (Barrick et al. 2001). First, Costa and McCrae (1992) describe conscientious humans as achievement striving, and Witt et al. (2002) suggest that humans with a high degree of conscientiousness work more thoroughly. Moreover, humans with high conscientiousness expression tend to be responsible, reliable, ambitious, and dependable (Costa and McCrae 1992). Barrick and Mount (1991) argue that people with a strong sense of purpose, obligation, and persistence generally perform better in most jobs. Indeed, it is difficult to imagine a position where one can be careless, lazy, impulsive, and low achievement striving (i.e., low conscientiousness) (Barrick and Mount 1991).

The second important predictor for job performance and teamwork is neuroticism (Barrick et al. 2001; Tett et al. 1991). Individuals with high neuroticism tend to have poor interpersonal relationships (Lopes et al. 2003). In contrast, individuals with a low degree of neuroticism are less vulnerable to negative affect and have better emotional control. Non-neuroticism (i.e., emotional stability) is essential to the accomplishment of work tasks in many professions, as anxiety, hostility, personal insecurity, depression, and not likely to lead to high work performance (Barrick and Mount 1991). Consequently, companies strive to find employees who have a low level of neuroticism.

2.2 Algorithmic Hiring

For HR departments, the examination of applications is a repetitive and time-consuming activity, with the difficulty of evaluating each applicant with the same attention focus (Wilson and Daugherty 2018). Using algorithmic decision making, firms can review a large number of applicants automatically. Therefore, due to growing pools of applications and simultaneously limited time of recruiters (Leicht-Deobald et al. 2019), firms are increasingly using algorithmic decision-based selection tools, such as asynchronous video interviews or telephone interviews, with an algorithmic evaluation (Dahm and Dregger 2019; Lee and Baykal 2017; Brenner et al. 2016). These algorithmic decision tools are being increasingly applied before applicants are invited to participate in face-to-face interviews (Chamorro-Premuzic et al. 2016; van Esch et al. 2019). With sensor devices, such as cameras and microphones, the verbal and non-verbal behavior of humans is captured and analyzed by an algorithm (Langer et al. 2019). For example, candidates answer several questions via video or telephone (Precire 2020), which are analyzed algorithmically. During the asynchronous video interview, candidates must record their answers to certain questions and upload them to a platform. Facial expressions (e.g., smiles, head gestures, facial expression), language (e.g., word counts, topic modeling, complexity, variety), and prosodic information (e.g., pitch, intonation, and pauses) are extracted by an algorithm, resulting in a personality profile of the applicant (Dahm and Dregger 2019; Naim et al. 2016). Previous studies have shown that faces and speech are rich sources of cues for predicting personality (Biel et al. 2012). Using modern technological advances, complete personal profiles, along with the FFM personality traits, are created.

Besides being time-efficient, the main objectives are to reduce the unconscious bias, enhance consistency in the decision processes, and seek fairer selection outcomes because human biases occur in in-person job interviews due to the human interpretation of answers (Lepri et al. 2018; Levashina et al. 2014). Grove et al. (2000) showed, in a meta-analysis, that mechanical prediction techniques are, on average, 10% more accurate than clinical predictions. In another meta-analysis of employee selection and academic admission decisions, Kuncel et al. (2013) found that the mechanical method's validity improves the job prediction by about 50 percent compared to a holistic data combination. Kuncel et al. (2013) also emphasized that experts even had more information than the algorithm in many cases but still made worse decisions (Grove et al. 2000; Kuncel et al. 2013). Facial and speech cues are the cues taken most into consideration when analyzing personality from a computational perspective.

Companies often argue that they implement algorithmic decision-making tools to prevent bias against certain groups and create a relatively fair selection process (Persson 2016). For example, Deloitte (2018) argues that the system processes each application with the same attention according to the same requirements and criteria. In a typical job interview, bias can occur when interviewers evaluate the applicant's non-job-related aspects, such as sex, age, gender, race, or attractiveness (Lepri et al. 2018; Levashina et al. 2014; Schmid Mast et al. 2011). Previous studies also revealed a biasing effect of physical attractiveness (Hosoda et al. 2003) and gender when considering an opposite-sex-type job (Davison and Burke 2000).

2.3 Fairness in Computer Science

Since fairness has been a central focus of interest for the longest time, ontological, psychological, and mathematical definitions of fairness exist (Lee and Baykal 2017). For example, Leventhal (1980) describes fairness as equal treatment based on people's performance and needs. With the expanding debate on algorithmic fairness (Dwork et al. 2012; Hardt et al. 2016), a plethora of fairness measures has been developed to quantify the fairness of the algorithm (Verma and Rubin 2018). Since the usage of machine learning algorithms and their validity and fairness is a topical problem, it is imperative to investigate the algorithm in case (Chouldechova and Roth 2018).

In the algorithmic fairness literature, the authors often focus on establishing fairness, either as pre- or post-classification alterations (Calmon et al. 2017; Hardt et al. 2016) or by using regularizations in classification problems (Kamishima et al. 2012; Zafar et al. 2015). Typical pre-classification alterations include re-weighing data (Kamiran and Calders 2012) or changing individual data points (Hajian and Domingo-Ferrer 2013). More recently, statistical frameworks for the pre-mapping of points have been published, which usually modify the data's estimated probability density to a fair representation (Zemel et al. 2013; Calmon et al. 2017). So far, these methods have been applied to binary classifications (Zemel et al. 2013) but are, in theory, extendable (Calmon et al. 2017).

Most modifications are proposed to achieve a deal with modifications in the modeling portion of the procedure. Here, constraint-based optimization, where constraints are based on individual notions of fairness, have been proposed (Kamishima et al. 2012; Zafar et al. 2015). However, note that all of these notions of fairness cannot be fulfilled simultaneously (Friedler et al. 2016). Furthermore, the trade-off between fairness and accuracy is discussed (Feldman et al. 2015), which shows that only adapting an algorithm during the modeling phase is most likely not worthwhile for stakeholders. Apart from these concepts, several contributions deal with measuring or detecting (un-)fairness, often as a post-process procedure (Kamishima et al. 2012). While theoretical contributions focus on measuring bias and de-biasing data sets, practitioners need domain-specific approaches and methodologies to automatically audit machine learning models for bias (Holstein et al. 2019). In the following, we will introduce several fairness measures to compare their usability in a recruiting context. These measures may be used as pre- or post-process measures and have partially influenced constraint-based optimization rules.

2.3.1 Fairness as the "80% Rule"

According to a guideline of the Equal Employment Opportunity Commission (EEOC v. Sambo's of Georgia 1981), the employment rates of one group should not be less than 80% of other group rates (Barocas and Selbst 2016). This "80% rule" has been picked up in current fairness literature and has been formalized for marginbased classification problems (e.g., by Zafar et al. (2015)). In a slightly simplified version, Friedler et al. (2019) have defined *disparate impact* as a division of probabilities of estimations for different groups. Specifically, let Y be a binary random variable to be predicted (such as the risk for recidivism or credit-worthiness) and $\hat{Y}$ its estimation. Furthermore, let there be a random variable G describing the group membership of a certain person in a certain group (such as “non-white” for $G = 1$ and “white” for $G = 0$). Then, disparate impact may be formalized as

$$DI := \frac{P(\hat{Y} = 1|G = 1)}{P(\hat{Y} = 1|G = 0)} \leq 0.8.$$

Further variants for the measurement of disparate impact include the measure by Calders and Verwer (2010), which, instead of a multiplicative comparison, calculates the difference between the conditional probabilities.

Another example of a fairness measure for binary classification is the comparison of false positive/negative rates motivated by the *equalized odds* definition of fairness (Friedler et al. 2019; Hardt et al. 2016; Verma and Rubin 2018). By implementation of Friedler et al. (2019), we define a fairness measure, equal opportunity via false negatives (EqOppoFN), as the ratio between the false-negative rates of different groups, where the variables $Y, \hat{Y}$ and G are defined as before:

$$EqOppoFN = \frac{P(\hat{Y} = 0|Y = 1, G = 1)}{P(\hat{Y} = 0|Y = 1, G = 0)} \geq 1.25$$

Note that, similar to the usage of the “80% rule” in the DI measure of disparate impact, we have chosen a distance of 80% between the false-negative rates, as 1/1.25 = 0.8.

Using these fairness measures, we will examine the effects of imbalances of the training data set of the algorithm with respect to, for example, gender or ethnicity. We are specifically interested in determining the answer to the following question: Despite high accuracy, does an unequal distribution of groups in the data set of an algorithm lead to an over- or underestimation of the likelihood to be invited for a job interview?

2.3.2 *Fairness as the Differences in Accuracy*

Furthermore, a large proportion of fairness measures detects differences in accuracies between protected and unprotected groups (Chouldechova and Roth 2018; Feldman et al. 2015; Friedler et al. 2019). In many cases, accuracy measures were developed for binary classification, such as the balanced classification rate

$$BCR = \frac{P(\hat{Y} = 1|Y = 1) + P(\hat{Y} = 0|Y = 0)}{2},$$

where Y denotes the true class of a data point and $\hat{Y}$ denotes the predicted class of a data point. The resulting fairness measure quantifies the difference in the balanced classification rates and was introduced by Friedler et al. (2019).

To extend the measures of fairness that were formulated for binary classification, we would like to evaluate the differences of two common accuracy measures as a way to define fairness measures for continuous variables. The first is the mean squared error (MSE), a classical tool for the evaluation of algorithms. The second one uses the mutual information (see, e.g., Cover and Thomas (1991)), which may be defined as

$$MI(\hat{Y}, Y) = \sum_{(\hat{y},y)\in D} P_{\hat{Y},Y}(\hat{y}, y) ln \frac{P_{\hat{Y},Y}(\hat{y}, y)}{P_{\hat{Y}}(\hat{y}) P_Y(y)},$$

with a normalization term, leading to

$$NMI(\hat{Y}; Y) = MI(\hat{Y}; Y)\sqrt{H(\hat{Y})H(Y)},$$

where H is an entropy function (also used by Strehl and Ghosh (2002)). The corresponding fairness measure is the difference between the normalized mutual information of the different groups.

2.3.3 *Class Imbalance Versus Representation Imbalance*

Regarding empirical findings on misrepresentations of data in a machine learning context, most of the publications deal with the so-called class imbalance problem (Al Najada and Zhu 2014), where classifications of imbalanced classes are made.

This type of imbalance deals with imbalanced classes in the desired output variable of the classification algorithm. As shown in Fig. 1, an unweighted support vector machine

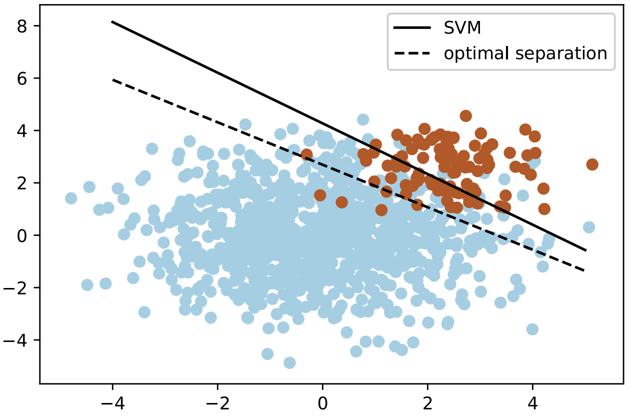

Fig. 1 Visualization of the class imbalance problem. Blue and red dots represent imbalanced classes. In this scenario, an unweighted support vector machine algorithm would lead to a suboptimal hyperplane and separation of the classes (color figure online)

(SVM) would overestimate the overrepresented classes and lead to an impaired performance. In contrast to that, imbalances in the representation of the data have been discussed in the context of worse performances of algorithms for certain subgroups of the population (Sapiezynski et al. 2017). The representation imbalance is often mistaken for the class imbalance problem, though it is conceptually different. The representation imbalance is visualized in Fig. 2, where a separating hyperplane may, for example, be optimal for the overall population while systematically disadvantaging certain subgroups of the population. Note that in Fig. 2, the classes are balanced. In fact, the number of blue and red dots is equal. Furthermore, balancing out the imbalanced groups would worsen the accuracy (drastically) instead of improving it.

While representation imbalance itself has not been extensively studied, there have been attempts of correcting biases. The calibration methods and correction of biases are mostly implemented as part of the modeling process (Feldman et al. 2015) and are restricted to a handful of algorithms (such as logistic regression (LR), SVM, naive bayes (NB)). Furthermore, random up- and down-sampling attempts, which are commonly used for class imbalance problems, will most likely impair the accuracy.

3 Method

To proceed with the criterion evaluation, we will introduce the data set and algorithms analyzed in this paper and the fairness measures used to evaluate them.

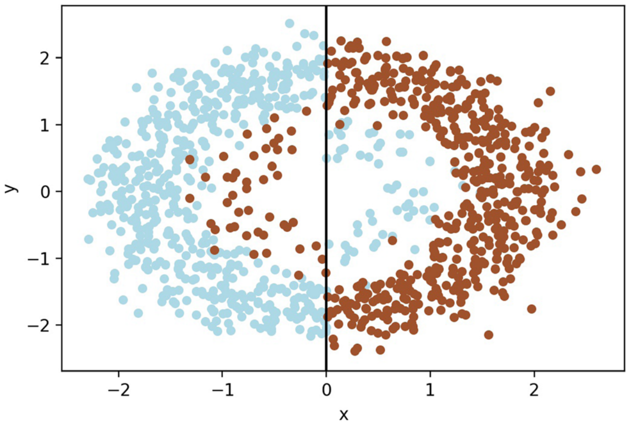

Fig. 2 Visualization of the representation imbalance problem. Blue and red dots represent two classes. The outer and inner rings represent two different groups of the population. The separating hyperplane assigns left to blue and right to red. For the outer (overrepresented) circle, and for the overall population, this separation is optimal, while it is the worst case for the inner ring (color figure online)

3.1 Description of the Data Set

ChaLearn is a non-profit organization hosting academic data science challenges, among which was the ChaLearn Looking at People 2016 First Impressions challenge intending to evaluate personality traits from YouTube videos (Ponce-López et al. 2016). Amazon Mechanical Turk (AMT) workers labeled these videos through a ranking procedure, which resulted in five-factor personality scores in the interval [0,1]. A year later, a second version ("V2") of this data set was released, with an extension of the data set introducing a job interview variable, which quantified the likelihood of the person in the video to receive a job interview invitation. Even though the experimental decision makers (AMT workers) were asked to make an invite-for-interview decision, the videos are not from a recruiting context but reflect content typically found on YouTube (i.e., beauty tutorials).

The First Impressions V2 data set, used in at least two challenges (Escalante et al. 2018; Ponce-López et al. 2016), contains 10,000 15-second videos collected from YouTube high definition (HD) videos and annotated with the help of AMT workers. These videos were extracted from over 3,000 different YouTube videos of people standing in front of a camera and speaking in English (Escalante et al. 2018). In each video, the person talks to a camera in a self-presentation context similar to video-conference interviews (Ponce-López et al. 2016). The participants are of different ages, gender, nationality, and ethnic backgrounds. The majority of videos are from Q&A and related contexts (e.g., vlogging, How-To´s, and beauty tips). In general, few humans appear in the video, and one unique person is in the foreground with a safe distance to the camera; and this person speaks with a clear voice and without much movement (Ponce-López et al. 2016). The number of videos from one channel was limited to three videos per YouTube channel. The videos' origin is quite diverse regarding views and 5-star ratings (Ponce-López et al. 2016). For the individuals in the videos, five-factor personality traits, as well as an "invite for interview" score (we will call this variable "job interview" score), were calculated utilizing the Bradley-Terry-Luce model (Bradley and Terry 1952). This resulted in five personality scores and a job interview score, each between 0 and 1, reflecting the degree of agreement with the given characteristic. Specifically, 0 represents the lowest possible agreement, and 1 indicates the highest level of agreement.

3.1.1 Structure of the Data Set

Since the data set was given within the framework of a challenge, parts were made available at different times. The construction of machine learning algorithms is often to predict data (test data) using previously collected data (training data). However, if using the test data in the algorithm's optimization process, this will implicitly lead to optimization of the algorithm on the test set and not on previously unseen data. In the case of unknown data, the true accuracy of the model might differ from the accuracy on the test set. To avoid this type of overfitting, the data set was split into three subsets as recommended, e.g., by Murphy (2012) as follows:

Training, validation, and test data sets: The *training data* set contains the video data (n = 6000), as well as ground truth annotations for the five-factor personality traits and the job interview variable. This data set was provided at the beginning of the challenge to train the algorithms. In the *validation data set*, video data (n = 2000) without annotations were given. Participants of the challenge were able to receive immediate feedback on their classification's performance on the CodaLab platform (Ponce-López et al. 2016). The unlabeled *test data* (n = 2000) were made available 1 week before the end of the challenge. The accuracy of this data set, measured as the MSE, determined their final scores. After the end of the challenge, all of this data and the labels were made available.

3.1.2 Biases in the Data

To analyze biases in the data, we used the annotations of gender and ethnicity by Escalante et al. (2018). These annotations were not used in any of the winning algorithms of the challenge. While the number of females and males appearing in the videos was somewhat balanced, Fig. 3 shows that the videos mostly depicted people of Caucasian ethnicity. In contrast, there was an underrepresentation of Asian and African-American ethnicities.

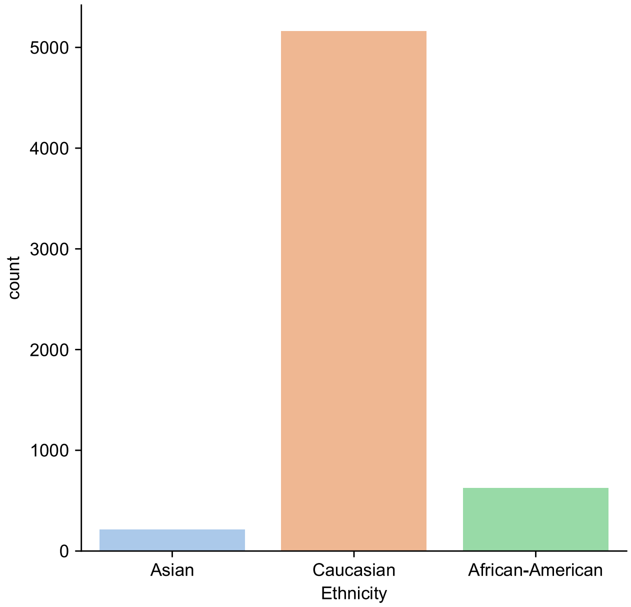

Fig. 3 Total number of Asians, Caucasians, and African-Americans in videos of the training data set (n = 6000)

Ponce-López et al. (2016) analyzed the First Impressions data set with descriptive methods and found several biases. They have also noted that, even though the same video was often segmented into 15-s fragments, there was a rather high intra-video variation of the labels. Because of the labeling procedure, which was mainly a ranking of all the videos, we would expect 50% as a fair mean for each of the groups' scores. Therefore, we mostly concentrated on deviations of the scores from 50%. When considering the means, there was a clear difference in the job interview score for males and females—females are slightly but significantly more likely to be invited for a job interview and had higher assigned values for conscientiousness and non-neuroticism (Ponce-López et al. 2016). As for ethnicity, Asians, in comparison to Caucasians, were more likely to be invited for a job interview. Table 1 also shows that there was a tendency of disfavoring African-Americans. These tendencies were equally apparent for the other traits (conscientiousness and non-neuroticism). In this sense, this data is comparable to real-world data: It reflects the biases and stereotypes existing in society, for example, a significantly higher level of conscientiousness among females (Goodwin and Gotlib 2004) and disfavoring African-Americans (Ford et al. 2004; Bertrand and Mullainathan 2004; Watson et al. 2011).

3.1.3 Limitations of the Data Set

The data set of the First Impressions V2 challenge has several limitations. First, personality traits are valid predictors for job performance, but whether one performs well in the job depends on the occupation and the situation (Tett et al. 1991). Due to different job demands, it is essential to consider the kind of job to accurately assess whether a person is suitable for this job. In the First Impression data set, the AMT workers were only told that they are human resource specialists who should select candidates for interviews (Tett et al. 1991).

Second, another limitation is that the videos do not originate from the context of recruitment, specifically job interviews; they are excerpts from publicly available YouTube videos. The videos are self-presentation videos from different Q&A settings, such as beauty or styling videos (Ponce-López et al. 2016). The limitation is that the behavior is probably slightly different in a job interview. However, since the videos come from a self-presentation context, the videos are similar to job interviews (one unique person in the foreground, presentation context, clear voice, little movement) (Ponce-López et al. 2016).

Moreover, the setting is similar to an asynchronous interview because applicants will try to present themselves in the best way, use impression management, and apply self-presentation strategies to convey a message or image (Chen 2016; Ma 2017). YouTubers make use of verbal expressions, nonverbal cues, and purposive behaviors (Ma 2017). The manner in which vloggers introduce themselves is similar to the interviewee's self-introduction to a recruiter (Ma 2017). YouTubers want to make an impression on their followers, which is comparable to the situation of a candidate who wishes to convince the recruiter.

Third, another limitation of the data set is that the raters were AMT workers without further qualifications or recruitment background. However, since a large number of videos have been evaluated and the AMT workers represent a cross-section of American society (Paolacci et al. 2010), and several experiments showed that AMT is an excellent opportunity to gain a representative sample of participants, e.g., Thomas and Clifford (2017), the results still can be considered as meaningful. Additionally, every person has a first impression of another person, even an experienced recruiter (Dougherty et al. 1994). This study is primarily concerned with how fair the algorithms reproduce the training data set. We assume that biases occur even in a professional setting and would therefore like to refine our research question to the reproduction of these biases.

3.2 Winning Algorithms

In the following, we will briefly introduce two of the top algorithms of the ChaLearn First Impressions V2 Challenge.

Model 1 (BU-NKU) Salah and colleagues (researchers from Bogazici University and Namik Kemal University in Turkey) submitted the algorithm with the best performance, measured as the smallest MSE when compared to the test set data (Kaya et al. 2017). At the feature level, they used face, scene, and audio modalities.

The first preprocessing step is recognizing facial features, where 49 landmarks are detected on each frame of a given video (Escalante et al. 2018). After cropping, resizing, and aligning the faces, features are extracted using the pre-trained VGG network (Parkhi et al. 2015). This system is then fine-tuned with respect to emotions using over 30,000 training images of the FER-2013 data set (Goodfellow et al. 2013). After the extraction of frame-level features, the videos are summarized using functional statistics, such as mean, standard deviation, and curvature (Escalante et al. 2018). These deep facial features are then combined with Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP), which applies Gabor filters on aligned facial images (Almaev and Valstar 2013). Furthermore, scene and acoustic features were extracted using the VGG-VD-19 network (Simonyan and Zisserman 2014), as well as an open-source tool called openSMILE (Eyben et al. 2010).

The modeling procedure involved an improved method for the choice of weights in single hidden-layer feedforward neural networks called extreme learning machine (ELM; (Huang et al. 2004)) together with a regularization coefficient for increased robustness and generalization capability (Escalante et al. 2018). The multi-modal ELM models are then stacked to a Random Forest (RF), an ensemble of decision trees, and programmed mostly in MATLAB (Kaya et al. 2017).

Model 2 (ROCHCI) Another submission came from the University of Rochester's Human–Computer Interaction department (ROCHCI). At the feature level, they used facial and audio modalities and the transcription of what was said in the videos. Four groups of features were used for the classification. The first group of features was hand-picked and -tuned from a facial tracker (available on GitHub: https://github.com/go2chayan/FacialAction). This involved, e.g., the position of the eyes and other landmarks (12 in total). Apart from this, they also used a tool called Praat to extract audio features, such as loudness or pitch of the sound. Furthermore, facial and meta attributes were extracted using SHORE, a commercially available face recognition software. As the last group of features, the video's transcription was used to implement simple word statistics, such as the number of unique words and the number of filler words. All of the features were then concatenated. The data was modeled using gradient boosting regression (Hastie et al. 2009) and was programmed mostly in Python.

The reimplementation of the algorithms involved several difficulties, which will be discussed in the following. Several toolboxes (which have to be purchased separately) are used. In the case of this examination, pre-implemented functions from missing toolboxes had to be re-implemented. As for the ROCHCI algorithm, it was mostly programmed in Python 2.7, which had to be migrated to Python 3.7. Furthermore, running the algorithms entails large calculation times, as the data is very large (around 16 GB). Furthermore, most algorithms used external software, which was either commercial or free. This adds a risk factor in replicating the results, as this software may no longer be available. Because of these difficulties, it cannot be guaranteed that these algorithms were replicated one to one. However, it was verified that the algorithms still yielded very high accuracies (over 0.98), which is the main concern of the examination at hand.

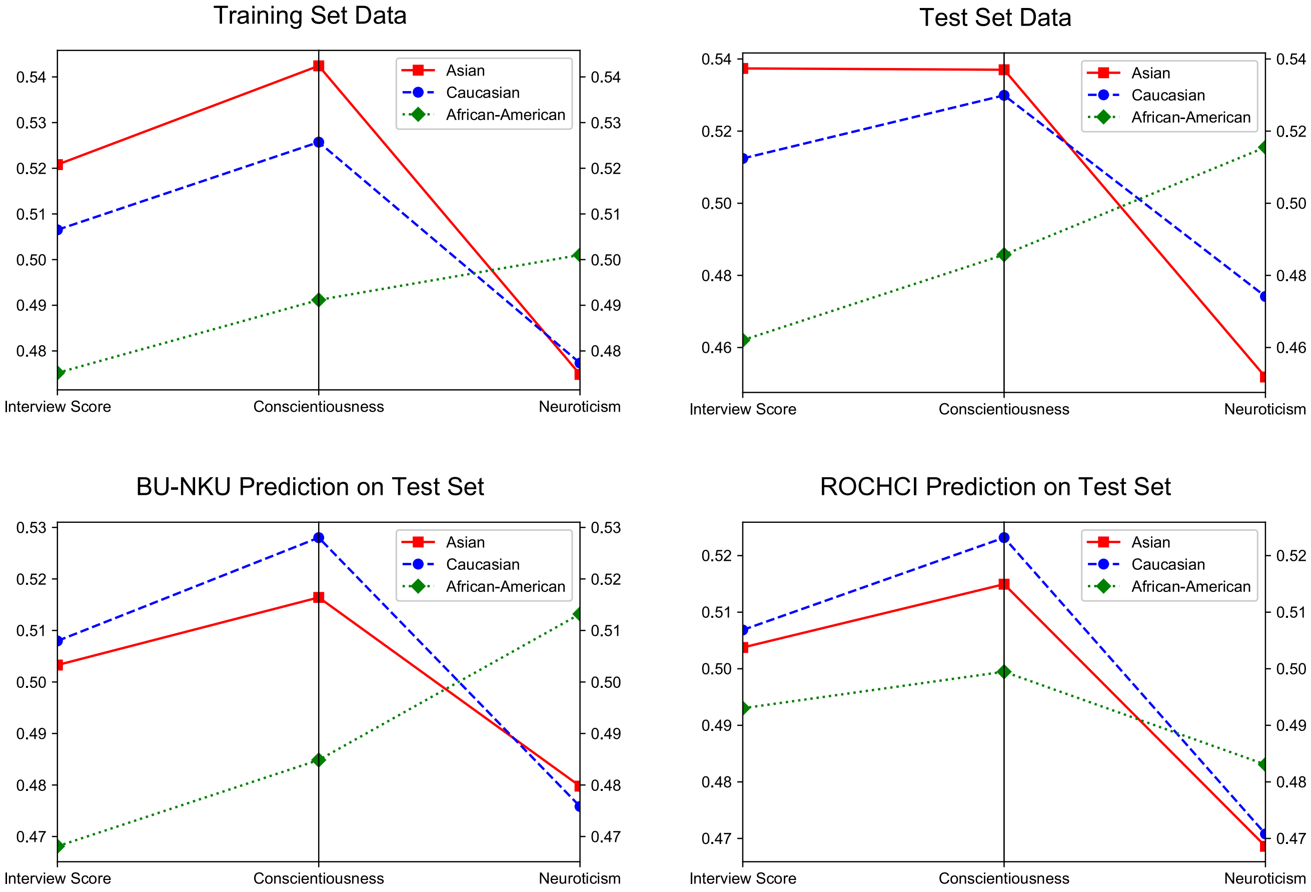

Fig. 4 Parallel plots of mean values for the job interview score, conscientiousness, and neuroticism scores in the training and test set, as well as the classifications of BU-NKU and ROCHCI on the test set

Table 2 Accuracy-based fairness measures applied to the job interview score in the classified labels of winning algorithms

Method	Male	Female	Asian	Caucasian	African-American
BU-NKU					
BCR	0.74	0.80	0.71	0.78	0.73
NMI	0.78	0.77	0.91	0.75	0.85
1-MSE	0.99	0.99	0.99	0.99	0.99
ROCHCI					
BCR	0.66	0.70	0.61	0.69	0.63
NMI	0.78	0.77	0.91	0.75	0.85
1-MSE	0.98	0.99	0.99	0.99	0.98

3.3 Connection to Practice

We aim to examine realistic scenarios and threats in connection to AI- and video-based automatic selection processes. Therefore, we conducted a thorough web-search for “off-the-shelf” products offering algorithms or environments for such an automatic selection process. We found 29 companies providing varying services. We contacted them personally and gathered information on their product’s technical details from their own or related websites. A list of the companies is available upon request. We discovered significant similarities between the different algorithms, such as the usage of transfer learning, usage of external products, usage of audio, video, and scenic data. Most surprisingly, one company, in fact, used the First Impressions data set for pre-training their own method. For the remaining companies, at least 15 used video data of applicants, four used external software or data, 12 used deep learning methods, and four used other types of machine learning algorithms. Only one company used a purely theory-driven method from psychological research. Note that 11 companies provided little or even no information about their product’s technical aspects on their website. In summary, we found significant similarities between “off the shelf” algorithms actually used in the industry and our reimplementation of the highly accurate winning algorithms.

Table 3 Group-comparison-based fairness measures applied to the job interview score in training and test set, as well as the classified labels of winning algorithms on the test set

Method	Male	Female	Asian	Caucasian	African-American
Training data set					
DI	0.97	1.03	1.01	1.03	0.96
Test set					
DI	0.89	1.12	1.18	1.25	0.72
BU-NKU (test set)					
DI	0.86	1.17	0.92	1.39	0.68
$\frac{1}{EqOppoFN}$	0.65	1.54	0.76	1.74	0.82
ROCHCI (test set)					
DI	0.82	1.22	1.05	1.15	0.83
$\frac{1}{EqOppoFN}$	0.64	1.56	0.95	1.34	0.80

Table 4 Statistics for the data classified by the BU-NKU and ROCHCI algorithms: *p*-values for paired t-tests and Cohen's d values for different attributes and ethnicities

Attribute	Group	BU-NKU		ROCHCI	
		p	d	*p*	d
Conscientiousness		0.01	− 0.38	0.01	− 0.40
Job interview score	Asian	0.13	− 0.23	0.12	− 0.26
Neuroticism		0.03	0.28	0.22	0.18
Conscientiousness		0.07	− 0.03	0.06	− 0.05
Job interview score	Caucasian	0.45	− 0.01	0.03	− 0.06
Neuroticism		0.50	0.01	0.29	− 0.03
Conscientiousness		0.36	0.05	0.00	0.30
Job interview score	African-American	0.90	− 0.01	0.13	0.13
Neuroticism		0.74	− 0.02	0.00	0.30

4 Results

We have re-implemented two winning algorithms of the ChaLearn First Impressions V2 challenge and evaluated the classifications in terms of fairness, and will present our results in the following.

4.1 Comparison of Means and Standard Deviations

The First Impressions V2 data set contains biases as well as an imbalanced representation of ethnicities. In Table 1, we have summarized the descriptive statistics of the training and test sets, as well as the classified labels of the winning algorithms grouped by ethnicities. In the training and test sets, Asian people were often preferred for job invitations compared to Caucasian people, and they, in turn, were preferred compared to African-American people. These preferential tendencies were similar for conscientiousness and non-neuroticism.

However, running the BU-NKU and the ROCHCI algorithms on the test set, Fig. 4 shows that the job interview score and conscientiousness are underestimated for Asian people. Asian people were given, on average, lower job interview scores than Caucasian people, with smaller standard deviations. Furthermore, the strength of the bias in favor of Asians differs between training and test set. It is striking that, even though Asian people had higher job interview scores in the training *and* the test set (see Table 1), the BU-NKU and ROCHCI algorithms strongly underestimated these values in their classifications.

Furthermore, as depicted in Table 1, the variances of all predicted job interview scores are smaller in comparison to the ground truth labels.

4.2 Accuracy and Fairness Measures

To evaluate the different algorithms, we began by computing different accuracy measures, as depicted in Table 2. We see that especially the measures NMI and BCR, which are also used in the fairness literature, seem to pick up the differences between the groups more sensitively compared to the MSE, which was optimized by the algorithms. In an attempt to quantify both the independence of the groups (Dwork et al. 2012), as well as the independence of the groups *given the outcome* (Hardt et al. 2016), we calculated the fairness measures DI and EqOppoFN. The results reported in Table 3 show that both of these measures pick up the biases found in a prior descriptive analysis of the data (see Table 1). For these measures, two extrema could be detected. First, as visualized in Table 3, the requirement of the DI value being above 0.8, i.e., the 80% rule, was not achieved for the group of African-American people when the BU-NKU classification was used. Furthermore, the ratio of negative classifications, as quantified by the EqOppoFN value, was the highest for the male group.

As can be seen in Table 3, neither the training nor the test set contained large (enough) group differences to result in a DI score below 0.8 for males. However, in the classified data, they fail the 80% rule multiple times.

Using paired t-tests, the significances in the differences between algorithmic scoring and true annotations of the test set were calculated, grouped by ethnicities and gender. The detailed results, when grouped by ethnicity, are depicted in Table 4. Here, the BU-NKU predictions showed significant discrepancies (with medium effect sizes) for Asians, and the ROCHCI predictions showed significant discrepancies for Asians and African-Americans. Regarding the job interview score, the ROCHCI predictions were significantly different for Asians and African-Americans, with effect sizes of − 0.3 and − 0.4, respectively. However, for Caucasians, the largest group in both data sets, we found no differences.

5 Discussion

We have replicated two highly accurate algorithms for classification tasks in the recruiting context and found that these algorithms still have deficits concerning inherent biases and unpredictable classifications. First, the predictive models replicate the bias existing in the training data. Both of the reproduced machine learning models transferred bias from the data set, even though they used very different algorithms (neural networks and decision trees, gradient boosting).

Second, consistent biases in the training data tend to be amplified by the predictive models when there is a balanced representation of groups in the data set. Both of the technically different, high-accuracy machine learning models increased the gender bias (favoring women) from the gender-balanced data set. As a result, the model's output failed the 80% rule test multiple times, even though the training data did not. The amplification of the bias does not impact the model accuracy between groups when measured by the MSE. As for the other fairness measures, the NMI seems to be less sensitive to tendencies and biases than, for example, the BCR.

And third, biases favoring or disfavoring underrepresented groups in the data set (in this case, ethnicity) may be both over- or underestimated by machine learning models. For example, in the case of ethnicities, Asians were disfavored even though this tendency was neither observable in the training data nor the test data; thus, this outcome was unpredictable. Unlike standard procedures in statistics, general guidelines or precise calculations for recommendable sample sizes do not exist in the machine learning literature. And much less for subgroup sizes, as in the case of Asians. Note that a small subgroup size may be permissible due to general contextual information, which can be extracted from any sample and is valid for the whole population (such as landmarks of the face, language, and so on). Therefore, we cannot exclude that this might be a sampling effect due to the small subgroup sample of Asians. Because of the known tendencies of machine learning algorithms in class imbalance, it may be natural to expect the tendency of over- and underestimation to transfer to the representation imbalance problem. However, we would like to point out that a rigorous proof of this has yet to be published to the best of the authors' knowledge.

This paper aimed at raising awareness about the possible difficulties regarding the unfairness of algorithmic decision making despite the high accuracy of the algorithm in the context of HR analytics. Previous research highlighted the advantages, such as cost and time savings (Suen et al. 2019; Leicht-Deobald et al. 2019), but knowledge of the potential problems of algorithmic decision making is still limited in the HRM literature. Using naturally occurring and realistic data, our findings add to the current knowledge in several ways. First, although companies stress the importance of implementing algorithmic decision making to become more objective and fairer in their recruitment process (Deloitte 2018), our results show that algorithmic decision making does not eliminate the threat of implicit biases and unfairness towards certain groups of people. Therefore, algorithms still lead to biased outcomes concerning gender, ethnicity, and personality traits if they build upon inaccurate, biased, unrepresentative, or unbalanced training data (Mehrabi et al. 2019; Barocas and Selbst 2016). In this case, the algorithm replicates and reinforces the existing biases and subjective prejudices in a society (Crawford and Schultz 2014).

Second, both algorithms replicated the biases of human judgments and (partially) amplified them. Females, for example, had higher job interview scores than males. Thus, even though algorithmic decision making should help companies to increase the objectivity and fairness of their recruitment process (He 2018), algorithmic decision making is not a panacea for eliminating biases, especially if the training data are inaccurate or unrepresentative in several ways. Complicating this issue, the specific kind of bias might be less apparent, as is the case for our data sets of the First Impressions challenge. For example, the ethnicity of the person in the video was coded after the challenge, and the resulting bias only appears because we tested the algorithm for these additional characteristics. Therefore, companies' recruiting functions need to know more about the specific aspects of the training data set used by service providers. Otherwise, there is a threat of excluding well-fitting candidates by the algorithm due to hidden biases.

Third, we found that underrepresentation in the data set might lead to unpredictable classifications. For example, there was an underrepresentation of Asians in the data sets used here, and in turn, both of the algorithms underestimated the job interview score for Asians. All applicants should have equal hiring opportunities, although underrepresentation in the applied algorithm's training data set reduces one's chance to get invited to a job interview if a person belongs to an underrepresented group. Therefore, when implementing algorithmic decision making, companies need to control and understand the training data set and should try to avoid any underrepresentation of certain groups of people or personal characteristics (Holstein et al. 2019). Otherwise, companies might jeopardize a diverse workforce in the enterprise, which is often a business priority (Economist 2019).

6 Practical Implications

There are important practical implications that follow from our results. First of all, our analysis shows that HR managers have to be careful when implementing algorithmic decision-based interview tools.

Our findings are in line with the notion forwarded by other researchers (e.g., (Langer et al. 2019; Holstein et al. 2019)) that companies must be cautioned to enforce and apply such algorithmic decision-making procedures carefully. Moreover, when implementing an algorithm, responsibilities and accountability must be clarified (Tambe et al. 2019). The HR management should cooperate with members of the organizations who have adequate expertise and a sophisticated understanding of the used tools to meet the challenges that the implementation of algorithmic decision making might face (Barocas and Selbst 2016; Cheng and Hackett 2019; Canhoto and Clear 2020). HR managers need to understand, with the help of the company's data scientists, how the algorithms operate (e.g., how the algorithm uses data and evaluates specific criteria) and disclose the aspects for the algorithmic decision. This comes with responsibility; organizations should clearly define humans responsible for applying algorithmic decision-tools (Lepri et al. 2018).

Furthermore, companies need to control the training data set and be responsible for applying the algorithmic decision-making tool (Lepri et al. 2018). Firms should implement proactive auditing methods (Holstein et al. 2019) since it is important to verify and audit the algorithmic decision process regularly (Kim 2016). Since fairness and, conversely, unfairness depends on the specific context (Lee 2018), and these contexts may vary remarkably, there is a need to develop automated auditing tools and innovative approaches to assess the context-specific fairness of algorithmic decision-making tools and machine learning (ML) systems (Holstein et al. 2019).

Firms investing in external service providers for HR algorithms need to know more about the training data set to evaluate if they (mis-)fit their company context. Often, the algorithm's code and training data set are not transparent to the clients (Raghavan et al. 2020; Sánchez-Monedero et al. 2020). For example, if a service provider trained its algorithms only on a specific ethnic group, the recruitment of international applicants might be biased if companies solely rely on the algorithm's suggestions for their hiring decisions. HireVue, for example, does not give detailed information about the training data set on their website. HireVue mentions that they do not have a one-size-fits-all algorithm. The data set of Precire consists of only 5201 persons representing the German population, including people with a German speech level of at least C1 (Stulle 2018). No information is available about the origin of the 5201 people (Linnenbürger et al. 2018). HR managers should receive detailed information about the data sets, the codes, and the service provider's procedures and measures to prevent biases. This information should be discussed with the company's data scientist because the interplay of domain knowledge and programming is indispensable.

In summary, companies should not rely solely on the information provided by algorithms or even implement automatic decision making without any human control. As a prominent example, Amazon's hiring algorithm yielded an extreme bias in favor of only male applicants, which finally led Amazon to shut down the complete automatic decision-making systems for their hiring decisions (Sackmann 2018). While a gender bias of algorithmic decision making in the case of Amazon seems obvious, implicit biases of less apparent characteristics might be more problematic because they are more difficult to identify (and test for).

7 Limitations and Future Research

The data set of the First Impressions V2 challenge has several limitations, as mentioned in Chapter 3.1.3 and in the study itself. The limitations give rise to further ideas for future research. First, it would be interesting to test video clips of interviews originating from the recruitment context. However, it is incredibly difficult to obtain videos from the recruitment context due to data protection regulations.

Moreover, the employment description must be more precise, and several types of jobs should be examined. In a further study, it would be interesting to test algorithms from the application context for their unfairness potential. We tried to obtain algorithms from the recruitment context (see Sect. 3.3), but most service providers were unwilling to share their algorithms for research purposes. A list of contacted service providers is available upon request.

Furthermore, we have restricted ourselves to reproducing highly accurate algorithms, as they might be used in the industry, instead of proposing ways to correct the bias. The ways of reweighting the data set or using constraint-based optimization in current research have been proposed for relatively simple data sets. It is unclear how to modify videos or the various feature engineering procedures for fairness. This also goes beyond the scope of this paper. It would have been possible to recalibrate the data after the procedures (which would mean to artificially "boost" the scores of, e.g., African-Americans by adding a small constant), but this would not only lack explainability but also be unrealistic for real-world purposes. Additionally, since a single definition of or a consensus on how fairness may be quantified does not yet exist (Mehrabi et al. 2019), the ways to achieve universal fairness remain unclear. The future research question could be: "To what extent can algorithmic de-biasing strategies be applied?" or "What are possible ways to avoid bias?".

Another future research avenue is to take a closer look at the difference in reliability and validity between algorithmic decision making and humans (Suen et al. 2019). Even though deviances from optimal fairness can be shown in many algorithmic decision-making settings, it is important, especially for practitioners, to know whether it would still be a positive change compared to human raters. Consequently, another future research question could be: "What is the difference in reliability and validity between AI-decision-makers and human raters?".

Furthermore, we only considered asynchronous video interviews and tested them for fairness. The literature is still at the beginning concerning other selection tools and assessing their fairness or unfairness, for example, gamification or algorithmic CV screening. Relatedly, a fruitful research avenue is a search for and detection of unfairness in other real-life data sets over and beyond observable characteristics (Mehrabi et al. 2019). Consequently, a complete algorithmic-based selection process with several stages could be tested for fairness or unfairness in future research.

8 Conclusion

In this paper, we show that even highly accurate algorithms can be discriminatory, and we highlight the ethical issues that might occur when using algorithms in the HR context. Our analysis emphasizes the importance of considering fairness aspects when implementing algorithmic decision making in the HR context. This article contributes to a better understanding of the unfairness potential of algorithms in HR recruitment. Companies are increasingly using algorithmic decision making in recruitment to save costs and achieve more objectivity in the recruitment and selection context. However, the utilization of algorithms in recruitment does not necessarily free companies from prejudices. As our study shows, the algorithmic outcomes can be biased, existing inequalities can be amplified depending on the training data, and unpredictable classifications can result from underrepresentation in the training data set. Therefore, it is essential to audit the quality of the training data set to prevent unfairness in advance. If companies use an algorithm in the hiring process, they risk losing well-fitting applicants because the algorithm does not put all suitable candidates on the shortlist for a job interview. In this case, the pre-selection of the algorithm is problematic, and the human recruiter is unable to detect or solve this issue.

Acknowledgment Open Access funding enabled and organized by Projekt DEAL. We gratefully acknowledge financial support from the Federal Ministry of Education and Research 01JD1812A/B.

References

Al Najada H, Zhu X (2014) iSRD: spam review detection with imbalanced data distributions. In: Proceedings of the 2014 IEEE 15th international conference on information reuse and integration. IEEE, Redwood City, pp 553–560

Almaev TR, Valstar MF (2013) Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. 2013 Humaine association conference on affective computing and intelligent interaction. IEEE, Geneva, pp 356–361

Anderson CW (1960) The relation between speaking times and decision in the employment interview. J Appl Psychol 44(4):267

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104:671

Barrick MR, Mount MK (1991) The big five personality dimensions and job performance: a meta-analysis. Person Psychol 44(1):1–26

Barrick MR, Mount MK, Judge TA (2001) Personality and performance at the beginning of the new millennium: What do we know and where do we go next? Int J Sel Assess 9(1–2):9–30

Barrick MR, Swider BW, Stewart GL (2010) Initial evaluations in the interview: relationships with subsequent interviewer evaluations and employment offers. J Appl Psychol 95(6):1163

Behling O (1998) Employee selection: Will intelligence and conscientiousness do the job? Acad Manag Perspect 12(1):77–86

Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Am Econ Rev 94(4):991–1013

Biel J-I, Teijeiro-Mosquera L, Gatica-Perez D (2012) Facetube: predicting personality from facial expressions of emotion in online conversational video. In: Proceedings of the 14th ACM international conference on Multimodal interaction, Santa Monica, pp 53–56

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. Biom 39(3/4):324–345

Brenner FS, Ortner TM, Fay D (2016) Asynchronous video interviewing as a new technology in personnel selection: the applicant's point of view. Front Psychol 7:863

Calders T, Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. Data Min Knowl Discov 21(2):277–292

Calders T, Žliobaitė I (2013) Why unbiased computational processes can lead to discriminative decision procedures. Discrimination and privacy in the information society. Springer, Heidelberg, pp 43–57

Calmon F, Wei D, Vinzamuri B, Ramamurthy KN, Varshney KR (2017) Optimized pre-processing for discrimination prevention. In: Advances in neural information processing systems 30, Long Beach, pp 3992–4001

Canhoto AI, Clear F (2020) Artificial intelligence and machine learning as business tools: a framework for diagnosing value destruction potential. Bus Horiz 63(2):183–193

Chamorro-Premuzic T, Winsborough D, Sherman RA, Hogan R (2016) New talent signals: Shiny new objects or a brave new world? Ind Organ Psychol 9(3):621–640

Chen C-P (2016) Forming digital self and parasocial relationships on YouTube. J Consum Cult 16(1):232–254

Cheng MM, Hackett RD (2019) A critical review of algorithms in HRM: definition, theory, and practice. Hum Resour Manag Rev:100698. https://doi.org/10.1016/j.hrmr.2019.100698

Chouldechova A, Roth A (2018) The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810

Costa PT, McCrae RR (1992) Four ways five factors are basic. Person Individ Diff 13(6):653–665

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Crawford K, Schultz J (2014) Big data and due process: toward a framework to redress predictive privacy harms. BCL Rev 55:93

Dahm M, Dregger A (2019) Der Einsatz von künstlicher Intelligenz im HR: Die Wirkung und Förderung der Akzeptanz von KI-basierten Recruiting-Tools bei potenziellen Nutzern. Arbeitswelten der Zukunft. Springer, Heidelberg, pp 249–271

Daugherty PR, Wilson HJ (2018) Human + machine: reimagining work in the age of AI. Harvard Business Press, Boston

Davison HK, Burke MJ (2000) Sex discrimination in simulated employment contexts: a meta-analytic investigation. J Vocat Behav 56(2):225–248

Deloitte (2018) Mensch bleibt Mensch - auch mit Algorithmen im Recruiting. Wo der Einsatz von Algorithmen hilfreich ist und wo nicht. https://www2.deloitte.com/de/de/pages/careers/articles/algorithmen-im-recruiting-prozess.html. Accessed 12 Sep 2019

Diakopoulos N (2015) Algorithmic accountability: journalistic investigation of computational power structures. Digit J 3(3):398–415

Dougherty TW, Turban DB, Callender JC (1994) Confirming first impressions in the employment interview: a field study of interviewer behavior. J Appl Psychol 79(5):659

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, (ACM), Cambridge, pp 214–226

Economist T (2019) How to make your firm more diverse and inclusive. https://www.economist.com/business/2019/11/07/how-to-make-your-firm-more-diverse-and-inclusive. Accessed 30 Nov 2019

Escalante HJ, Kaya H, Salah AA, Escalera S, Gucluturk Y, Guclu U, Baró X, Guyon I, Junior JJ, Madadi M (2018) Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos. arXiv preprint arXiv:180200745

Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on multimedia, Firenze, pp 1459-1462

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, pp 259-268

Ford TE, Gambino F, Lee H, Mayo E, Ferguson MA (2004) The role of accountability in suppressing managers' preinterview bias against African-American sales job applicants. J Pers Selling Sales Manag 24(2):113–124

Frieder RE, Van Iddekinge CH, Raymark PH (2016) How quickly do interviewers reach decisions? An examination of interviewers' decision-making time across applicants. J Occup Organ Psychol 89(2):223–248

Friedler SA, Scheidegger C, Venkatasubramanian S (2016) On the (im) possibility of fairness. arXiv preprint arXiv:160907236

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, Atlanta, pp 329-338

Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Trans Inf Syst 14(3):330–347

Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee D-H (2013) Challenges in representation learning: a report on three machine learning contests. International conference on neural information processing. Springer, Heidelberg, pp 117–124

Goodwin RD, Gotlib IH (2004) Gender differences in depression: the role of personality factors. Psych Res 126(2):135–142

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical versus mechanical prediction: a meta-analysis. Psychol Assess 12(1):19

Hajian S, Domingo-Ferrer J (2013) Direct and indirect discrimination prevention methods. Discrimination and privacy in the information society. Springer, Heidelberg, pp 241–254

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Conference on neural information processing systems (NIPS), Barcelona, pp 3315-3323

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Heidelberg

He E (2018) Can artificial intelligence make work more human? Strateg HR Rev 17(5):263–264

Holstein K, Wortman Vaughan J, Daumé III H, Dudik M, Wallach H (2019) Improving fairness in machine learning systems: What do industry practitioners need? In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1-16

Hosoda M, Stone-Romero EF, Coats G (2003) The effects of physical attractiveness on job-related outcomes: a meta-analysis of experimental studies. Person Psychol 56(2):431–462

Huang G-B, Zhu Q-Y, Siew C-K (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In: International joint conference on neural networks (IEEE Cat. No. 04CH37541). IEEE, Budapest, pp 985-990

Huffcutt AI, Conway JM, Roth PL, Stone NJ (2001) Identification and meta-analytic assessment of psychological constructs measured in employment interviews. J Appl Psychol 86(5):897

Hurtz GM, Donovan JJ (2000) Personality and job performance: the Big Five revisited. J Appl Psychol 85(6):869

Junior JCSJ, Güçlütürk Y, Pérez M, Güçlü U, Andujar C, Baró X, Escalante HJ, Guyon I, Van Gerven MA, Van Lier R (2019) First impressions: a survey on vision-based apparent personality trait analysis. IEEE Trans Affect Comput, p. 1-20. https://doi.org/10.1109/TAFFC.2019.2930058

Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33(1):1–33

Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. In: Joint European conference on machine learning and knowledge discovery in databases, Springer, pp 35–50

Kauermann G, Kuechenhoff H (2010) Stichproben: Methoden und praktische Umsetzung mit R. Springer, Heidelberg

Kaya H, Gurpinar F, Ali Salah A (2017) Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, Honolulu, pp 1-9

Kim PT (2016) Data-driven discrimination at work. Wm & Mary Law Rev 58:857

Kuncel NR, Klieger DM, Connelly BS, Ones DS (2013) Mechanical versus clinical data combination in selection and admissions decisions: a meta-analysis. J Appl Psychol 98(6):1060

Langer M, König CJ, Papathanasiou M (2019) Highly automated job interviews: acceptance under the influence of stakes. Int J Sel Assess. https://doi.org/10.1111/ijsa.12246

Lee MK (2018) Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc 5(1):2053951718756684

Lee MK, Baykal S (2017) Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. ACM, Portland, pp 1035-1048

Leicht-Deobald U, Busch T, Schank C, Weibel A, Schafheitle S, Wildhaber I, Kasper G (2019) The challenges of algorithm-based hr decision-making for personal integrity. J Bus Ethics 160(2):377–392

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 31(4):611–627

Levashina J, Hartwell CJ, Morgeson FP, Campion MA (2014) The structured employment interview: narrative and quantitative review of the research literature. Person Psychol 67(1):241–293

Leventhal GS (1980) What should be done with equity theory? In: Gergen KJ et al (eds) Social exchange. Springer, Boston, pp 27–55

Lindebaum D, Vesa M, den Hond F (2019) Insights from the machine stops to better understand rational assumptions in algorithmic decision-making and its implications for organizations. Acad Manag Rev 45(1):247–263. https://doi.org/10.5465/amr.2018.0181

Linnenbürger A, Greb C, Gratzel DC (2018) PRECIRE technologies. Psychologische Diagnostik durch Sprachanalyse. Springer, Heidelberg, pp 23–56

Lopes PN, Salovey P, Straus R (2003) Emotional intelligence, personality, and the perceived quality of social relationships. Pers Individ Differ 35(3):641–658

Ma X (2017) How May I Impress You? A content analysis of online impression management tactics of YouTube Beauty Vloggers. Master thesis, University of Nevada. https://digitalscholarship.unlv.edu/thesesdissertations/3090/. Accessed 28 Oct 2020

Marler JH, Boudreau JW (2017) An evidence-based review of HR analytics. Int J Hum Resour Manag 28(1):3–26

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2019) A survey on bias and fairness in machine learning. arXiv preprint arXiv:190809635

Möhlmann M, Zalmanson L (2017) Hands on the wheel: Navigating algorithmic management and Uber drivers'. In: Proceedings of the international conference on information systems (ICIS), Seoul, pp 10-13

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press, Cambridge

Naim I, Tanveer MI, Gildea D, Hoque ME (2016) Automated analysis and prediction of job interview performance. IEEE Trans Affect Comput 9(2):191–204

Paolacci G, Chandler J, Ipeirotis PG (2010) Running experiments on Amazon Mechanical Turk. Judgm Decis Mak 5(5):411–419

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Proceedings of the British machine vision conference (BMVC), Swansea, pp. 41.1-41.12

Persson A (2016) Implicit bias in predictive data profiling within recruitments. In: IFIP international summer school on privacy and identity management. Springer, pp 212-230

Ponce-López V, Chen B, Oliu M, Corneanu C, Clapés A, Guyon I, Baró X, Escalante HJ, Escalera S (2016) Chalearn lap 2016: First round challenge on first impressions-dataset and results. In: European conference on computer vision, Springer, pp 400-418

Precire (2020) Precire technologies. Precire technologies. https://precire.com/. Accessed 3 Jan 2020

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

Rothstein MG, Goffin RD (2006) The use of personality measures in personnel selection: What does current research support? Hum Resour Manag Rev 16(2):155–180

Sackmann C (2018) The doors of intelligence. https://www.focus.de/finanzen/boerse/die-tuecken-der-intelligenz-amazon-schaltet-ki-ab-die-bewerbungen-von-frauen-als-minderwertig-erachtete_id_9741890.html. Accessed 30 Oct 2019

Sánchez-Monedero J, Dencik L, Edwards L (2020) What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. ACM, Barcelona, pp 458-468

Sapiezynski P, Kassarnig V, Wilson C, Lehmann S, Mislove A (2017) Academic performance prediction in a gender-imbalanced environment. Proc FATREC Workshop Responsible Recomm 1:48–51

Schmid Mast M, Bangerter A, Bulliard C, Aerni G (2011) How accurate are recruiters' first impressions of applicants in employment interviews? Int J Sel Assess 19(2):198–208

Shankar S, Halpern Y, Breck E, Atwood J, Wilson J, Sculley D (2017) No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:171108536

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556

Springbett B (1958) Factors affecting the final decision in the employment interview. Can J Psychol 12(1):13

Strehl A, Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3(Dec):583-617

Stulle KP (2018) Psychologische Diagnostik durch Sprachanalyse: Validierung der PRECIRE®-Technologie für die Personalarbeit. Springer, Heidelberg

Suen H-Y, Chen MY-C, Lu S-H (2019) Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? Comput Hum Behav 98:93–101

Suresh H, Guttag JV (2019) A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:190110002

Tambe P, Cappelli P, Yakubovich V (2019) Artificial intelligence in human resources management: challenges and a path forward. Calif Manag Rev 61(4):15–42

Tett RP, Jackson DN, Rothstein M (1991) Personality measures as predictors of job performance: a meta-analytic review. Person Psychol 44(4):703–742

Thomas KA, Clifford S (2017) Validity and Mechanical Turk: an assessment of exclusion methods and interactive experiments. Comput Hum Behav 77:184–197

van Esch P, Black JS, Ferolie J (2019) Marketing AI recruitment: the next phase in job application and selection. Comput Hum Behav 90:215–222

Verma S, Rubin J (2018) Fairness definitions explained. 2018 IEEE/ACM international workshop on software fairness (FairWare). IEEE, Gothenburg, pp 1–7

Vinciarelli A, Mohammadi G (2014) A survey of personality computing. IEEE Transact Affect Comput 5(3):273–291

Watson S, Appiah O, Thornton CG (2011) The effect of name on pre-interview impressions and occupational stereotypes: the case of black sales job applicants. J Appl Soc Psychol 41(10):2405–2420

Wilson HJ, Daugherty PR (2018) Collaborative intelligence: humans and AI are joining forces. Harvard Bus Rev 96(4):114–123

Witt L, Burke LA, Barrick MR, Mount MK (2002) The interactive effects of conscientiousness and agreeableness on job performance. J Appl Psychol 87(1):164

Zafar MB, Valera I, Rodriguez MG, Gummadi KP (2015) Fairness constraints: mechanisms for fair classification. arXiv preprint arXiv:150705259

Zehlike M, Hacker P, Wiedemann E (2020) Matching code and law: achieving algorithmic fairness with optimal transport. Data Min Knowl Discov 34(1):163–200

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: International conference on machine learning, Atlanta, pp 325-333