Machine Learning and Gender Bias in Accounting Job Recruitment

Sheilla Njoto (s.njoto@unimelb.edu.au)

[1]Ly Fie Sugianto (lyfie.sugianto@monash.edu)

Leah Ruppanner (leah.ruppanner@unimelb.edu.au)

Abstract

This paper discusses the impact of biases in Machine Learning and how it could lead to unintended consequences. It is an attempt to set a perspective on the importance of data and algorithm when employing Machine Learning in predictive analytics. The paper provides an overview on Data Analytics and Machine Learning as the emerging topics in today's information era. In particular, it describes what is meant by Machine Learning and how Machine Learning can be utilized for automated decision making in the Accounting discipline. It also depicts a case on the use of Machine Learning for Accounting job recruitment in Australia and in the United States. Our results show that our female candidates are disadvantaged in two ways: (1) the Machine Learning algorithm viewed gender specific name more favorable; (2) human recruiters were more likely to view and call the male applicant for interview. We demonstrate two clear pathways through which female job applicants experience occupational discrimination in job recruitment process.

Introduction

Prior to the era of Big Data, accountants work with structured numeric data in the form of spreadsheets and tabulated data. With recent technology advances in capturing, storing, processing and analysing large data sets, there has been a transformation in the way accountants work with data. Big Data & Analytics (BDA) together with Machine Learning (ML) and Artificial Intelligence (AI) enable data driven analysis and automation of massive data sets, including social media archives and financial reports in answering important business questions. The era of Big Data brings about the important change in the accounting profession, such as utilising ML, unstructured data and visualization tools. Today, we have seen the applications of BDA with ML and AI in Accounting and Finance (see Brown, Crowley & Elliott, 2020, Jiang & Jones, 2018, Purda & Skillicorn, 2015; Mayew & Venkatachalam, 2012; Holton, 2009).

[1] Corresponding author – Ly Fie Sugianto is an Associate Professor at the Department of Accounting, Monash University.

Accountancy today is becoming less transaction driven and more value focused. It is imperative that professionals working in this discipline need to keep up with this trend. While these changes are taking place progressively in the workplace, there is an urgent need for educators to develop learning material to help prepare future accountants with the right skill set. As endorsed by the Association to Advance Collegiate Schools of Business (AACSB), business schools are required to deliver curriculum content that cultivates agility with current and emerging technologies (Standard 4). Accounting students need to be equipped with analytics mindset to possess the right skill set to cope with the demand of the future professions.

Accounting academics have begun responding to the challenge posed by Big Data & Analytics by integrating analytics into accounting curriculum to equip students with relevant skills for future workforce where automation and human-machine partnership become the norm. Accounting education literature today has reported different ways of infusing data analytics into the accounting curriculum. Borthick and Smeal (2020) introduced a data analytic case in taxation which can be used as part of a Taxation course (at both undergraduate and postgraduate levels as a small group project) or an IT Auditing course (at the postgraduate level as an individual assignment). Dzuranin, Jones & Olvera (2018) proposed three ways of implementation methods: (1) a focused approach– whereby the foundation of data analytics skill is provided in a stand-alone course; (2) an integrated approach – whereby data analytics content is infused into existing accounting courses; and (3) a hybrid approach whereby the accounting programs include both a stand-alone course emphasizing data analytic competencies and accounting courses with data analytics competencies ingrained. Further, based on a review of academics and professional literatures on applications, tools and potential limitations or consequences, Dzuranin et al. (2018) identified three focus areas for accounting analytics program, namely (1) the ability to ask the right questions, (2) the need to understand data and perform analysis, and (3) the ability to communicate the results of the analyses. While accounting educators tend to consider developing students' mindset as the most important aspect in the curriculum, the importance of data bias seems to be overlooked in the accounting analytics program. In fact, PwC in its 2015 report had emphasised the importance of data, including how non-traditional data, such as images or words, can be utilized leading to better insight. Further, it also recommended the ability to research and identify anomalies and risk factors in underlying data as the new, evolving skill set that needs to be possessed by accounting graduates (PwC, 2015).

Furthermore, the growing deployment of automated decision making in the accounting field is made possible by machine learning. The extent of automation and the interaction of human-machine

partnership in decision making is an emerging issue which needs attention, especially with the growing emphasis on effective, relevant and ethical decision making within the accounting profession. Regulatory bodies, such as Australian Human Rights Commission and Tertiary Education Quality and Standards Agency, have emphasized the importance of fairness and equal opportunities when organisations are using big data and utilising machine learning to facilitate decision making. Up to date, there is hardly any discussion in the accounting education literature on human-machine partnership in decision making and its potential drawback due to bias.

This paper intends to shed some light on algorithmic and data bias in machine learning and how it affects equal opportunities and diversity in the workplace. The remaining of the paper is structured as follows. The next section discusses Data Analytics and Machine Learning as the emerging topic in the accounting curricula. It presents an overview on how ML and automated decision making offers exciting possibilities to be applied in accounting and finance. It also introduces the concept of machine learning and what is meant by learning in this context. The subsequent section discusses the importance of data and some risks associated with ML. Then, a case on the use of ML in job recruitment in a gender balanced occupation is depicted. We conducted two experiments to explore issues surrounding equal opportunity by submitting job application documents to accounting recruitment bots. The result section reports our findings from the Australian context and the US context. This is followed by a discussion on the result, including the implications and recommendations on accounting curriculum, organization and professional practices and digital ethics. It is hoped that the paper will inform academics when preparing accounting students for a future workforce where human-machine partnerships become prevalent. It is also our hope that the paper will inform professional bodies, policy makers and managements to continually endorse values such as diversity and inclusion, to ensure fairness and equal opportunities in employment, education and other areas in life.

Machine Learning

Advances in our ability to capture, store, process and analyse large data sets are transforming the accounting landscape. With the growing availability of voluminous data, there has been an increasing demand for efficient method to derive insights from the data. Automated decision making in the form of predictive analytics has been widely accepted and used in the accounting field. Behind the predictive analytics is a class of algorithms known as Machine Learning.

In its broader definition, ML is a family of procedures that acquire learning from data sets, generate predictions based on the learning experience and improve at tasks based on the repeated training experience (Domingos, 2012; Raub, 2018). ML algorithms that depend on specific data sets undergo supervised learning. They acquired the learning from the training data set. Another class of ML which does not depend on training data set is known as unsupervised ML. For the sake of simplicity and keep the discussion focused, this paper highlights the application of supervised ML. A simplification of a supervised ML is depicted in **Figure 1**.

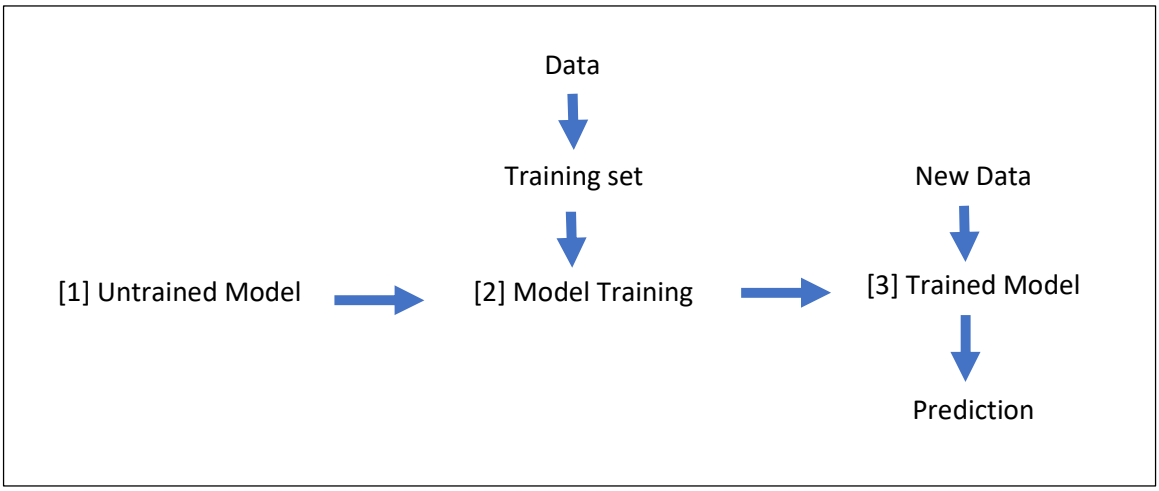


Figure 1: Model Creation Process (Source: Costa et al., 2020).

As shown in Figure 1, specific data sets are used to train ML. The datasets can be obtained from a variety of sources, depending on the chosen algorithm and its desired outputs. The chosen data set needs to undergo a preparation process known as Extraction-Transformation-and-Loading (ETL) to ensure the data is clean and ready to use as training data set. In the following stage, as specified in the second column of Figure 1, historical or existing data that essentially captures the predictive knowledge is labelled. For example, in the case whereby ML is trained to perform a classification problem, the data set may consist of a class of "successful" candidate and a class of "unsuccessful" candidates. Further, the data set is partitioned so that a portion of the data is used for training while the other portion is used for validating purpose. Model training is a process whereby a specific ML algorithm, such as Deep-Learning Neural Network (DNN) and Decision Tree, is employed to tune the algorithm into a trained model. When the model recapitulates the corresponding patterns in the training data, it creates a generalization. This is what is referred to as 'learning' in ML. This procedure manifests differently in different forms of algorithms—oftentimes categorized by the levels of human intervention in labelling and classifying the data.

Following this, the algorithmic model processes these datasets to analyze new data and create predictions. Note here that the term 'prediction' in this case does not necessarily refer to generating *future* predictions. Rather, it turns new data into classification (such as labelling spam emails), regression (such as generating quantitative risks of recidivism) or information retrieval (such as reducing selected findings on search engine). However, the realm of data analytics encompasses predictive analytics which deals with classification problem, such as the one described above, and prediction problem, which aims to predict a target value, such as credit score, fraud score or interest rate.

Notably, these predictive analytics are designed in different forms. For instance, modern ML techniques, such as the DNN, consist of more complex models that instinctively improve in accuracy as more data and cases are used. Contemporary critiques towards this center around the concerns towards its incomprehensibility due to its algorithmic complexity and therefore, its lack of mathematical proof of procedure (Barocas & Selbst, 2016; Boscoe, 2019; Kolkman, 2020; Miller, 2019 Mittelstadt *et al.*, 2016; Tolan, 2018). Miller (2017) proposes that complex models such as this lack explainability for two reasons: (1) the model structure is built upon numerous networks, statistical weights and therefore imitate what is termed as 'neurons' parallel to a human brain; and (2) the vast volume of training data used can oftentimes create mathematical idiosyncrasies that are inexplicable. These critiques are an integral part of this paper as it will later highlight the significance of biases that occur in predictions.

To understand the notion of learning in ML and comprehend the extent of learning capacity in ML, it is useful to contrast the concept of learning with human learning. Human learning can be described in many ways. When associated with cognitive ability, learning can be understood as the competency to acquire or deduce knowledge. At this higher level of functioning, cognitive learning is often associated with the competency to think, perceive, memorize, judge, reason and interpret. Evidently, this high-level functioning is important for decision making and planning. Hence, it is fair to conclude that human learning is almost always future looking as effective learning is a survival skill to cope with future situation and to make prediction. Furthermore, in the learning process, one may encounter uncertain and doubtful situation. In fact, in many cases, making mistakes can serve as valuable learning experience; and accordingly, a feedback mechanism is an integral part of learning whereby one can reflect the consequence of the chosen course of action or decision, and explore another course of action which leads to a better outcome. When the cognitive ability is working with psychomotor ability, learning can be understood as the ability to acquire new skill, associated with movement, coordination, strength and speed, such as in swimming, dancing and playing musical instrument. The outcome of such learning process can be observed more evidently, as often there are measurable changes associated with the learning process. In other words, performance outcome can be measured in terms of the acquired skill resulting from repetitive learning or training.

The cognitive ability in human is at a very high level that it cannot be modelled by programmers nor mimicked by machines. An individual learns by interacting with others, things, symbols and signs by exploring, experimenting, investigating and innovating. As learning progresses, the individual's cognitive system is further developed. Human capacity to extend knowledge is something far beyond the existing state of machine learning. Further, there are other levels of functioning associated with learning competency in human, such as the ability to mimic others – referred to as social learning. This, too, cannot be mimicked by machines.

In short, the notion of learning in ML can be best described as training with examples. The repetitive training is necessary for the entity, in this case referred to as the machine, to become aware of consequences, to commit into memory and to receive instructions. The change of behavior, as a result of the training process, can be observed in terms of performance improvement. With repetitive trainings using sufficient dataset, the machine can improve its accuracy in making prediction.

Data Analytics and Machine Learning in Accounting

Since the early 2000s, the value of computer-based automated decision making (ADM) has been recognised by administrators and decision-making bodies to provide a significantly efficient and accurate assistance (Cheong & Leins, 2020). For general business transactions, automation has been deployed to capture financial transactions and records these transactions in large databases. Traditional data entry routines can be automatically performed using optical sensor and character recognition software program, resulting in extraction of specific texts. Data extraction can also be automated by using scripting language to combine various data from different data repositories, including databases, for further analysis. Apart from process efficiency and effectiveness, the main impact of automation and data analytics in accounting is through facilitation of decision-making process, namely when the extracted data can be presented using visualization dashboard for auditing or other accounting purposes.

With automation, it is also possible to implement continuous auditing to raise alarm or exception when an auditing rule is violated. In this case, identifying alarms and exceptions can be modelled as a classification problem. In an auditing scenario, a notification of an alarm or exception will be attended by auditors who will follow a set of procedures to resolve the issue. Auditors must determine whether or not a transaction that raises the alarm is indeed problematic, such as a fraudulent transaction, or the alarm is false because the transaction can be considered normal. Identifying fraudulent claims associated with

payment in insurance setting and anomaly detection in payments for insurance beneficiary can be modelled as a classification problem.

Another commonly known application of machine learning is in targeting bank customers who are likely to respond to loan solicitations. In addition, ML can be used to classify credit rating systems when banks or financial institutions automate the decision to approve loan applications. Existing studies have applied hybrid models combining different ML techniques. Experiments using a real-world dataset from a bank in Taiwan were reported in by Tsai and Chen (2010).

In inventory management, ML can be used to predict inventory obsolescence. A classification model may be used to determine the probability of items which are obsolete or current. The same concept can be used for warranty management and prediction, resulting in companies avoiding warranty claims.

Biases in Machine Learning

Notwithstanding the significant impact of human-machine partnerships in decision-making, there is a growing literature critiquing the role of ML in exacerbating biases in the society (Chong & Leins, 2020; Costa *et al.*, 2020; O'Neil, 2016). In expounding this point, we acknowledge four forms of biases that can occur in ADM, as proposed by Kim (2019), namely intentional bias, record error bias, statistical bias and structural bias. In addition, we highlight data bias and distinguish it from record error bias.

Intentional bias

Intentional bias is the discriminatory action that is intentionally ingrained within the algorithms in order to explicitly box in or out some groups out of the equation (Kim, 2019). This can happen when ML designer intentionally includes (or excludes) certain characteristic(s) in (or out of) the model. Of course, companies can use this to reverse bias and to battle the existing structural inequality. For woman applicants, this would require a human intervention to tilt the algorithm to include them into their talent pools. However, note that intentional bias can lead to discrimination against women in conscious and unconscious ways.

Structural bias

Structural bias can be understood as the bias manifested by the system or the underlying phenomena captured in the model. Although the ML algorithm is able to capture the relationships

between the variables accurately, and that the model is not bias in statistical sense, the underlying characteristics of the model is inherently bias.

Record error bias

Record error bias refers to the bias that occurs due to incomplete or error in data. In job recruitment context, when ML makes a hiring prediction based on erroneous data, it may lead to an incorrect and subsequently unfair outcome as the applicant misses the employment opportunity.

Data bias

Data bias refers to the bias that occurs due to the use of misrepresentative data set. As an example, in 2014, Amazon generated hiring algorithms to predict the suitability of applicants for job positions. This system was trained using the internal workforce data recorded over the past 10 years (Costa *et al.* 2020). In 2018, it was found that Amazon's hiring algorithms discriminated against women and feminine language (Bogen 2019; Dastin 2018; O'Neil 2016). However, this discrimination was not intentional; rather, it was a consequence of the algorithm's training of biased data that captured existing inequality (Costa *et al.* 2020; O'Neil 2016). Specifically, as the majority of Amazon's employees were white men, it meant their hiring algorithms utilized this pattern as a determining factor of success, and therefore, discriminating against female candidates (Costa *et al.* 2020; Faragher 2019). Keywords such as "all-women's college" and "female" served as proxies that ranked female applicants lower (Costa *et al.* 2020; Faragher 2019). Amazon shut this system down soon after this algorithmic bias was found (Dastin 2018).

As evidenced by this case, data is hardly neutral (Bush, Lyons & Miller, 2020). Without a balanced representation of protected groups, algorithms may include demographic attributes as one of the variables for success predictions. Without careful mitigation, algorithms are prone to making correlations between success and attributes that do not serve as performance or qualification measures. This is not exclusive to gender-indicating keywords. In fact, some recruitment algorithms have been reported to make correlations between the keyword 'creativity' and their retention in a particular role (O'Neil, 2016). They can also make associations between the keyword 'inquisitive' and their likelihood of finding other opportunities and therefore lower retention. When these correlations are made based on demographic traits, such as neighbourhood, race or gender, the algorithms are at risk for bias. For women, this may lead to discrimination based on schooling, certain types of extracurricular activities, employment gaps for parental leave, and/or other correlated gender characteristics.

Statistical Bias

Statistical bias is closely linked to record error bias and data bias. It refers to the prediction of an individual based on the statistical representation of the demographic group into which this individual is categorised. This is exemplified by the 2016 case whereby a ProPublica article (Angwin *et al.,* 2016) shed a light on a recidivism risk assessment tool programmed by Nortpointe Inc., known as COMPAS. This system was criticised to introduce biases against African American defendants, ranking them with a higher risk score compared to their white American counterparts. ProPublica claimed that a defendant is rightly concerned with the probability that they will be incorrectly classified as high-risk, due to their ethnic background. This demonstrates that the algorithm mis-predicted the risk of an African American defendant as it drew statistical calculations based on racial representation of criminal records.

Another case study, exemplifying the interaction between statistical and data bias is the use of predictive-policing software (Byrne & Cheong 2017; Ferguson 2017). In a report, Ferguson (2017) identified that the modern application of predictive policing does not align with its own original intention; which they term Predictive Policing 1.0, using platforms such as *PredPol*. Here, instead of *preventing crime to property* in certain locations, the wrong assumptions are used in *predicting individuals' propensity to offend* within those locations. With these wrong assumptions, police departments will act by sending more resources to the area predicted to have more crime; hence more of that crime is reported; creating a positive feedback loop. In other words, this can be characterised as the model building becomes a self-fulfilling prophecy in entrenching prejudice against minorities – the more police resources deployed in an area, the more people of colour will be identified by police as suspicious, leading to a justification for prolonged police presence (Ferguson 2017; Byrne & Cheong 2017). This case study highlights how both the feedback loop above and *overfitting* the model to the data (not discounting the wrong assumptions in model building) – amongst others, covered in Ferguson (2017) - can have negative implications to civil liberties and worsens unjust treatment based on race, gender and class.

Without a comprehensive mitigation, algorithms are prone to statistical bias, data bias and structural bias. Its implication to gender bias is also immense and is not limited to explicit biases. To take the previous example, as hiring algorithms correlate 'creativity' to length of employment within the same job (O'Neil 2016), they can disadvantage women who are more likely to have career interruptions due to caregiving, lowering their probability rate for higher retention. In this respect, gender bias is not always explicit but measured through proxies programmed into the algorithms.

These cases serve as evidence that even without direct human interference, AI and recruitment can replicate existing biases (Bogen 2019). Individuals from historically discriminated groups are the most negatively affected by ADMs (Lambrecht & Tucker 2020). This paper, therefore, seeks to depict a case using the context of recruitment, both in Australia and in the United States. Here, we take one step in uncovering this process through a CV experiment that manipulates the gender. This allows us to identify whether the CV is seen differently based on gender-differences in applicant name alone.

The use of Machine Learning in Job Recruitment

In the context of this study, it is crucial to understand the context of recruitment in the past years and how ML situates in this setting. In the past decade, recruitment has been one of the vital components that determine the success of an organization (Kulkarni & Che, 2017). Economically, as applied to other aspects of corporate management, the efficiency of a talent acquisition is targeted to minimize cost for maximum corporate benefits (Kulkarni & Che 2017; Schweyer, 2010). This implies that recruitment is not limited to retention and hiring but encompasses an overall system that calculates the corporate loss and gain with every employee replacement. This is where the use of ML becomes practical. ADM efficiently sorts a massive volume of applications with minimum cost, but also is includes an inherent misconception that automation *should* give rise to impartiality (Kulkarni & Che 2017; Preuss 2017). Theoretically, AI *should* negate bias by creating an optimum amalgam of candidates ranked on merit. Unfortunately, as evidenced above, many real-world instances have shown that algorithms are no less prone to bias than humans (see O'Neil 2016; Costa *et al.,* 2020, etc.).

This is not to generalise all hiring algorithms nor their functionality. In fact, hiring algorithms, including their source of data, the level of human interventions and the types of predictive outputs, take different forms. Typically, though, different tools would perform a somewhat identical principle: trained algorithms screen new data and produce predictions based on a set of success metrics (Kulkarni & Che, 2017). The main difference, however, lies on the: (1) type of data, (2) algorithmic models, and (3) source of success metrics. The type of dataset is a major component that construes a prediction. While most recruitment software is limited to CV parsing, others may include additional components, such as automated interviews, background screening, and cognitive or psycho-tests. It thereby follows that some of the software uses a variety of data structures, including image or voice recognition, texts, or

big data (Van Esch *et al.,* 2019). Thus, it is essential to use rigorous empirical methods to scrutinize these algorithms to uncover bias.

For this study, we focus on CV parsing tools that are embedded within online recruitment tools, both in Australia and in the United States. These tools allow applicants to create their professional profiles, including their name, phone number, email, location, qualifications, professional and academic background and key skills. Applicants can search for job listings and apply using their accounts, attaching their CV and cover letter. On top of this, hirers can search for talents and offer job openings to them.

The Experiment

We apply an experimental design to exemplify how bias can occur in human-machine partnership decision-making using gender variables in recruitment setting. We conducted a CV experiment which has been well used across a range of studies (Costa *et al.* 2020; Kim 2019; Raub 2018). Our methodological approach is simple as our only manipulation is a gendered name change: applying using the same CV with one marked James (or Michael) and one Jane (or Sarah). Since the CVs are **exactly the same**, we have reduced the variation that is typical of a real CV. This allows us to explicitly isolate the role of gender in structuring human-machine decisions.

We applied to mid-level accounting jobs that were strategically selected because they are gender-balanced in Australia, which is marked 48.3% for men and 50.5% for women (ABS 2018) as well as the United States, which is marked 53.5% for women (US Labor Statistics, 2019). This means the demand for women's talent should be equal to that of men. Accounting positions had the additional benefit of being high on the list for recruiters which provided another data point for comparison. We developed mock CVs for a mid-career person (aged 35 years) and targeted jobs vacancies that require 7-10 years of prior experience. We developed career profiles with the top accounting companies and only applied to job postings that were full-time to ensure we were competing in the most competitive positions. We applied to 180 open positions in Australia and 50 open positions in the United States. Both experiments span over a 3-month time in 2020, conducted towards major CV-parsing online recruitment tools. These tools typically aggregate job vacancies from different companies and advertise them on an online basis. Note that one of the predominant functions of this online tool is to parse CVs and rank them based on suitability to the given job description. It offers a candidate management tool to

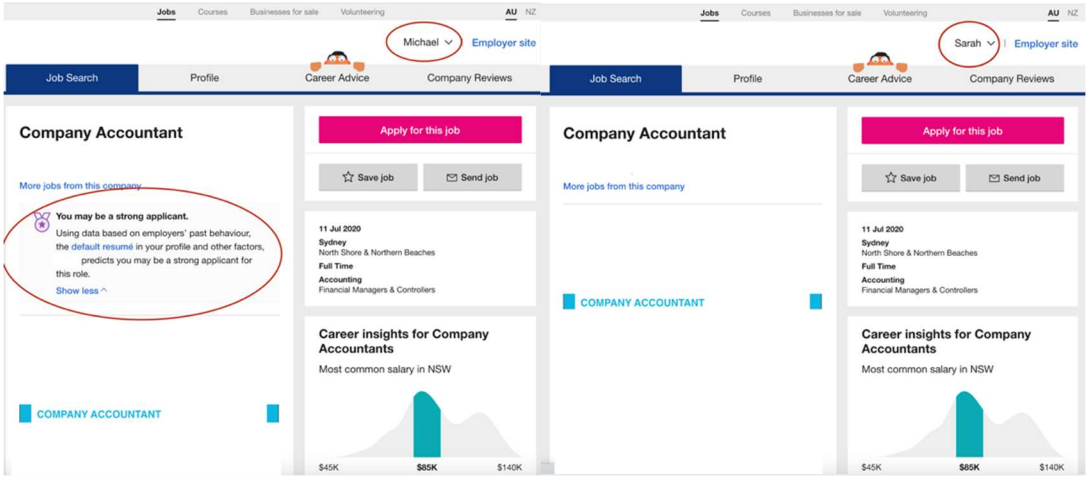
employers, wherein the algorithms embed ranking stars next to the applying candidate profile. Here, we introduce manipulations to human-machine recruitment decision-making.

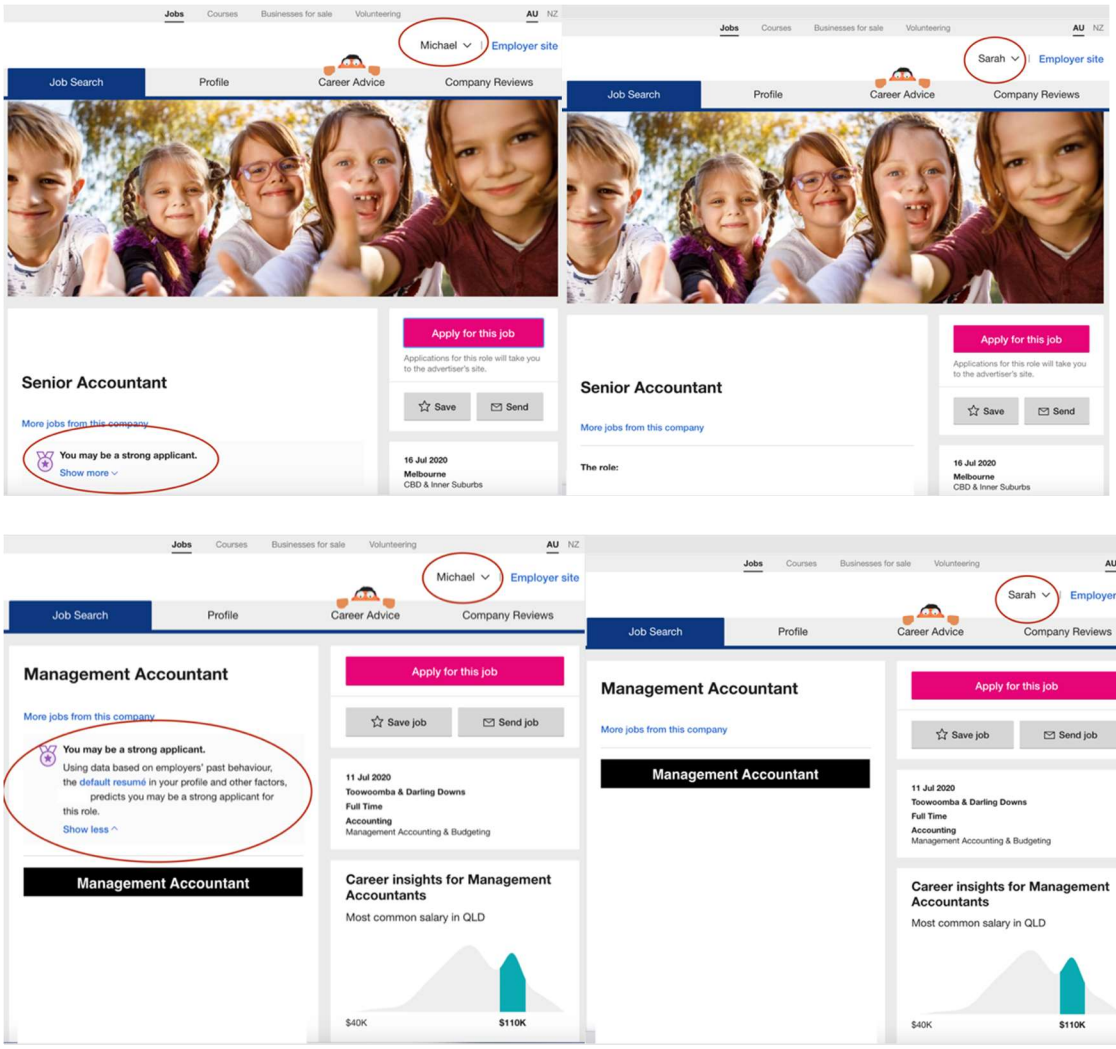
This process allowed us to address our two main research questions: (1) *do hiring algorithms discriminate against women in selecting candidates?;* and (2) *do we identify gender differences in the human interaction (profile looks, recruitment calls) for our candidates?*

The Result: the Australian context

Our initial question is whether we can document gender bias in a human-machine partnership hiring decision. We sent through two of the exact same CVs for the same job, one named Michael and one named Sarah. For two of the positions, we got a clear indication that the hiring algorithm was explicitly programmed to account for the gender of the applicants' name in sorting the applications. Figure 2 shows parallel outcomes for an equivalent job application for a senior accounting manager. When we submitted the CV with the name Michael, we received a badge that stated "You may be a strong applicant. Using data based on employers' past behaviour, the default resume in your profile and other factors, (we) predicts you may be a strong applicant for this role." The women's CV, which was equivalent except for name, did not receive an equivalent badge encouraging our applicant for this position.

Figure 2. ***The appearance of the male (badge) and female (without badge) profiles - Australia***





Because we cannot see the components that comprise the hiring algorithm used by this major search engine, we cannot determine whether the female candidate is being ranked lower because of employer preferences from applicants or the use of existing data to train the machine (machine learning). The fact that our male candidate did not receive a badge for each job indicates that the bias is not industry-wide but rather is utilizing some combination of employer and employee data that preferences our male candidate for these positions.

We also estimate how human recruitment may bias against women. We identify these data through two sources: (1) downloads of our resumes on the search engine app; and (2) texts or phone calls for job interviews or recruitment. Table 1 shows our male resume had 67 profile views compared to 30 views for our female CV. Our female resume received only 11 resume downloads, but our male resume was downloaded 41 times by prospective hirers.

Table 1. ***Results of the second stage experiment in Australia***

Subjects	***No. of profile views***	***No. of resumé downloads***
Michael (man)	67	41
Sarah (woman)	30	9

Table 2 shows the number of calls, texts, and emails for our woman and man candidates for the jobs. We see that our man candidate received more than twice as many positive calls or texts for the position (58 versus 23), three times as many positive emails and 9 fewer rejection emails than our woman candidate. When employers reached out to our fictitious applicants, our man candidate received positive contact 80.95% of the time compared to 52.45% for our woman candidate. Clearly, our man candidate was perceived more favourably than our woman candidate through the human contact with their CVs.

Table 2. ***Human Contact with CVs accessed through Search Engine***

Subjects	***No. of job applied***	***No. of positive calls/texts***	***No. of positive emails***	***No. of rejection emails***	***Total no. of responses***	***% of positive responses****
Michael (man)	180	58	27	20	105	80.95%
Sarah (woman)	180	23	9	29	61	52.45%

*in comparison to all responses (not total job applications).

The Result: the US context

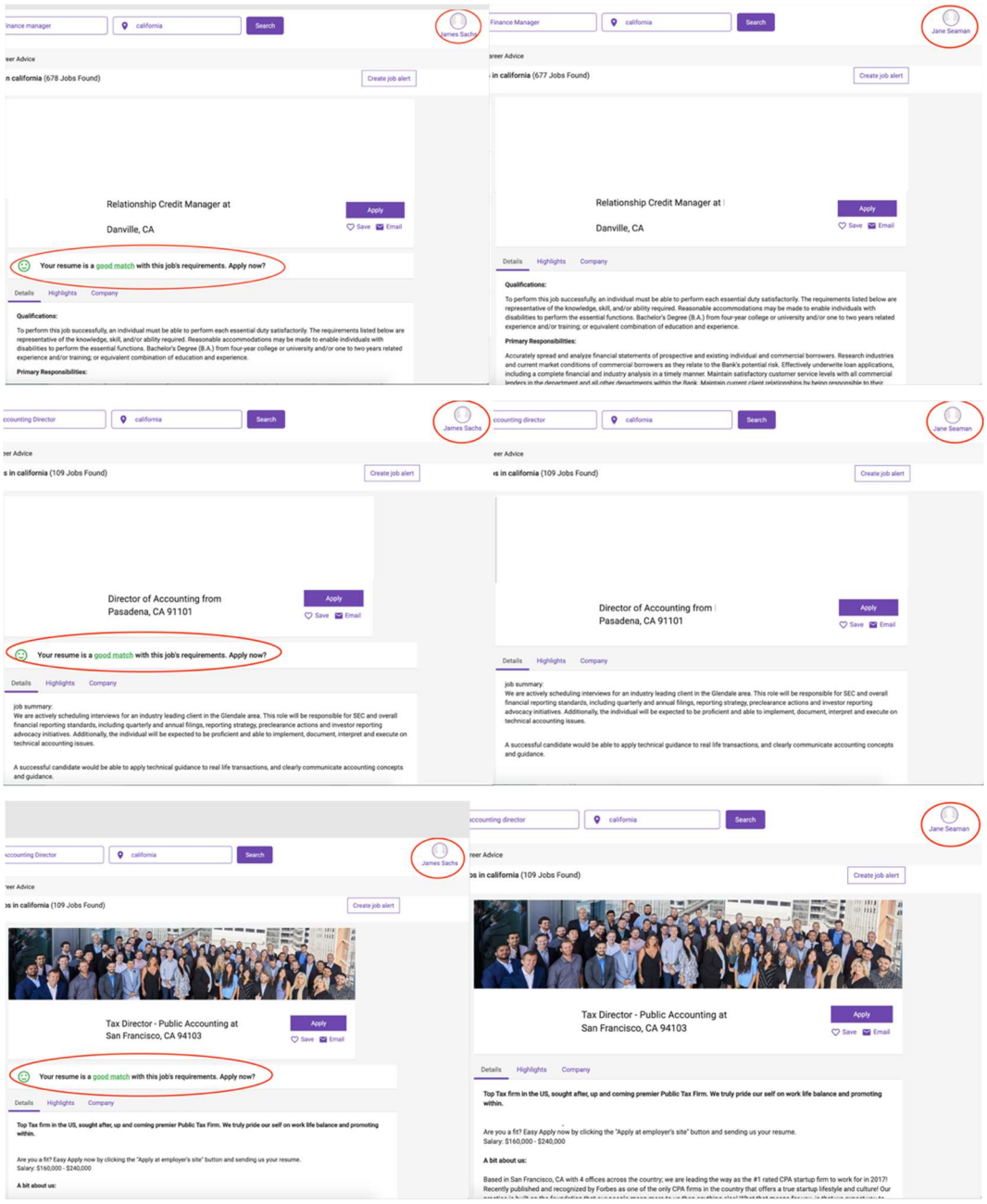
A similar pattern is also apparent in the US recruitment setting. For the same jobs, our man candidate was given an indication that his resumé was 'a good match' with (the) job's requirements.' Here, 'good match' indicates that the 'patented semantic matching technology evaluates (their) resumé to determine how well it matches with (the) job'. Out of 50 jobs we applied to, our man candidate received 27 'good match' badges, whilst our woman candidate received none. **Figure 3** below exemplifies the outlook of these badges on James' profile in juxtaposition to Jane's without the badge.

Table 3 demonstrates that there is a similarity of outcome in the US as that seen in the Australian context. Our man candidate received more 30 positive call-backs, in comparison to our

Table 3. ***Human Contact with CVs accessed through Search Engine***

Subjects	*No. of job applied*	*No. of cold call job offers*	*No. of positive emails*	*No. of positive calls*	*Total no. of responses*
James (man)	50	20	6	30	56
Jane (woman)	50	12	2	18	32

Figure 3. ***The appearance of the man (good match) and woman (without) profiles - US***



woman candidate with only 18 calls, with three time more positive emails. On top of that, our man candidate received almost double the job offer emails (aside from the positions we applied to), as our woman candidate did (20 versus 12).

Discussion

Our results suggest that human-machine partnership in recruitment fails to reduce gender bias, counter to expectations that it is more neutral. We use a case study of accounting jobs recruitment to document that with the same qualifications, experiences and word choices in CVs, the final outcome of recruitment favors men more than women. We find multiple points at which our women candidates are disadvantaged. First, the algorithm was more likely to encourage our man applicant to apply for the job by providing a badge indicating he was a good fit for the position. Our woman applicant did not receive equivalent encouragement despite having **the same CV** demonstrating that the gender of our applicant's name, rather than the characteristics of the CV.

Notably, the algorithms may vary within occupation developing specific decision-making based on an organizations' previous hiring decisions or interaction on the platform. Thus, women may experience greater algorithmic discrimination in positions whereby the humans interacting with the hiring algorithms lists exhibit preferences for men candidates. We cannot parse out these iterative processes in algorithms development for at least two reasons. First, the algorithms are often proprietary. Second and more importantly, the important capital of algorithmic predictions is not just data, but the networks between those data points. Predictions are typically generated from the correlational links between one point to another. For this reason, most computer scientists take pride in the state-of-the-art algorithmic models such as DNN, given its higher levels of complexity and automation, which often gives the impression that it produces much more accurate predictions. Algorithmic models such as this are much harder to audit, even by computer scientists. The programmers may very generally know how their algorithms would work but they are unable to explain how their algorithms produce the outputs. As evidenced by the findings, it is clear that the algorithms capture the correlation between success metrics and a candidate's gender. This is problematic because, first, in basic statistics, we understand that correlation does not mean causation; and second, because gender affects hiring predictions.

Yet, our selection of an occupation with gender parity in Australia and high women's representation in the US should mitigate some of this damage. However, more work applying our

experimental design across occupations with varying levels of gender representation is essential to understand the scope of the problem. Further, our selection of a typical Anglo-Saxon first name (James, Michael, Jane and Sarah) should reduce intersecting forms of race- and gender-based discrimination. Again, more research on these experiences by race are also necessary.

In addition to differences in algorithmic outcomes, we also found our man candidate was more likely to have his CV downloaded and be contacted after applying for the job. Thus, our man candidate also experienced a hiring premium from the human interaction with the hiring outcomes. This provides a clear and teachable example of how the human-machine interaction can impose gender bias in decision-making. The machine, using available data reflecting previous human hiring decisions and broader societal data, is more likely to rank the man higher but our human panel who interacts with the machine is also more likely to select the man candidate. Similar to previous research, this creates a feedback loop that has the potential to further reinforce the discrimination. For this reason, adjusting the algorithms to reduce gender bias is unlikely to completely alleviate the issue. Below, we offer some clear policy solutions and teaching recommendations with these complexities in mind.

Implications and Recommendations

In Curriculum and Teaching

This paper has outlined the proliferation of automated decision making enabled by Machine Learning in organizations today. In addition, the job recruitment case depicted in this paper clearly illustrates how algorithmic and data biases can lead to unintended consequences. Highlighting the issue of biases in ML has two educational implications.

First, it underlines the importance of teaching critical thinking to examine the mechanism used by Machine Learning beyond its training performance, its capability to predict accurately and the efficiency it offers. The algorithm employed in Machine Learning and the data used can be inherently biased. This potential problem should be acknowledged and addressed. ML should be trained with reliable and balanced data set. Algorithmic bias should be dealt with appropriately following cutting edge research and informed practice in Computer Science.

Second, it offers accounting students with an alternative pivot point to realize that the engagement of an automated, programmable solution should not be driven solely by business value. The role of human decision maker in the human-machine partnership in the current context is to ensure that

the decision made does not disadvantage a particular group or under-represented group in the society. Particularly, in the context of decision making involving human attributes, careful human intervention is needed to prevent the system to produce bias outcome, which in the long term leads to the creation of digital monocultures.

These two points can be regarded as educators' responsibilities to develop digital competencies and acumen beyond the standard learning material that is merely instructional in nature. Educators also need to shape students' worldview through an advocacy that cultivates the importance of digital ethics and promotes diversity and inclusion. Students should realize that both human and machine have limitations and our worldviews are often biased and incomplete. Recommendations from ML should be evaluated from different perspectives to avoid problematic outcome.

On Praxis

The findings through the experiments depicted in this paper reveal that bias still exists even in the seemingly gender-balanced profession. We offer two recommendations to recruitment companies which uses automated recruitment bots. First, we would like to highlight the importance of data policy, as poor data quality leads to poor decision making. As part of ML training process, results produced by ML should be analyzed to minimize, bias.

Second, we recommend that organizations with recruitment bots establish digital code of conduct to mitigate arising issues caused by algorithmic and data biases. Internal audit of information systems that include ML should be performed periodically. The digital code of conduct should be included as part of the terms and condition of recruitment bots. It can also be published as part of formal corporate ethics documents to indicate the corporate's commitment to equal opportunity. Professional bodies should expect the inclusion of digital code of conduct to be part of best practice for organizations.

On Human-centric Digital Ethics

It is no doubt that automated systems further humans' capability in decision-making and processing information. Although it is evident that it presents us with ethical concerns, its adoption in society should not be discouraged. The concern here is not incorporating ML into decision-making, but our emphasis on its merit and compromise on its social consequences and digital ethics and therefore our unreadiness to respond to them.

First, programmers and users need to shift their focus from pursuing efficiency in a broad term. We need to be critical of the objective of the system and pursue effective decisions without compromising

on human-centric values. In less than a decade, the scholarship of digital ethics has come up with diverse formulations of human-centric algorithmic principles—from fairness, transparency, accountability, privacy and more (Tsamados et al. 2021; Mittelstadt et al. 2016). Many even critique the value of complex algorithmic models such as DNN. Although they may offer advanced mathematical intricacy, it may not produce the most effective decisions for the targeted outputs. For example, the above hiring case shows us that the correlational points between success and gender is not necessarily the most effective decision for the targeted objective, although efficient. Recently, a few modern concepts emerged, suggesting that AI should be able to explain their own decisions. This includes explainable AI (Miller 2017) and interpretable AI (Rudin 2019; Rudin & Radin 2019; Zheng *et al.* 2016). Explainable AI refers to algorithms that can be explained in human terms; whilst interpretable AI refers to algorithms that explain their own decisions and elucidate the cause and effect of its mathematical weights. Of course, these models cannot eliminate biases within the system, but we are able to see when the systems introduce biases.

Second, we need to consider using ML not only to assist us with menial and systematic tasks, but also to help humans implement ethical standards, such as automating oversight, auditing conflicts of interest, or even incorporating non-tech equality frameworks such as gender quotas. This does not only help humans to mitigate biases in resolving tasks but also support humans to pursue fairness in doing so.

Lastly, we need to shift our focus from purely techno-centric perspective to a broader lens. Indeed, we are faced with technological challenges—however, we see from the above case that technologies do not necessarily introduce biases, but they capture the already existing societal problems within the data and perpetuate them, if not exacerbate them. Not only that we need technical fix to mitigate this issue, but we need to understand the broader outlook of the problem itself. For example, gender discrimination in hiring cannot only be fixed by gender-sensitive algorithms but it requires a deliberate systematic reform, such as reforming corporate culture, putting incentives for women to enter the workforce, educating recruiters on gender equality, and more. To pursue equality, we need to understand that inequality exists—and that is a concept that algorithms have yet to understand.

References

Australian Bureau of Statistics 2018, 'Census of Population and Housing: Reflecting Australia – Stories from the Census, 2016', *Australian Bureau of Statistics*, viewed 22 May 2020, <https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/2071.02016?OpenDocument>.

Bertrand, M and Mullainathan, S 2004, 'Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labour market discrimination', *American Economic Review*, vol. 94, no. 4, pp. 991–1013.

Bogen, M 2019, 'All the ways Hiring Algorithms Can Introduce Bias', *Harvard Business Review*, 6 May, https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias.

Borthick, AF and Smeal, LN 2020, 'Data Analytics in Tax Research: Analyzing Worker Agreements and Compensation Data to Distinguish Between Independent Contractors and Employees Using IRS Factors', *Issues in Accounting Education American Accounting Association*, vol. 35, no. 3, pp. 1—23.

Brown, NC, Crowley, RM and Elliott, WB 2020, 'What are you saying? Using Topic to detect financial misreporting', *Journal of Accounting Research*, vol. 58, no. 1, pp/ 237—291.

Cohen, S 1976, 'The basis of sex-bias in the job recruitment situation', *Human Resource Management*, vol. 15, no. 3, pp. 8–10.

Costa, A, Cheung, C and Langenkamp, M 2020, *Hiring Fairly in the Age of Algorithms*, Research Paper, Cornell University.

Cunningham, LM and Stein, SE 2018, 'Using Visualization Software in the Audit of Revenue Transactions to Identify Anomalies, *Issues in Accounting Education American Accounting Association*, vol. 33, no. 4, pp. 33—46.

Dalenberg, DJ 2018, 'Preventing discrimination in the automated targeting of job advertisements', *Computer Law & Security Review*, vol. 34, pp. 615–627.

Dastin, J 2018, 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*, 10 October, viewed 20 April 2020, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Dickson, B, et al. 2018, 'What is Algorithmic Bias?', *TechTalks*, 27 March, bdtechtalks.com/2018/03/26/racist-sexist-ai-deep-learning-algorithms/.

Dzuranin, AC, Jones, JR and Olvera, RM 2018, 'Infusing data analytics into the accounting curriculum: A framework and insights from faculty', *Journal of Accounting Education*, vol. 43, pp. 24—39.

Ernst & Young Foundation (2016), *Introduction to the analytics mindset*. Available at: <https://eyo-iis-pd.ey.com/ARC/ARC_default_XY/asp>.

Faragher, J 2019, 'Is AI the enemy of diversity?', *People Management*, 6 June, viewed 20 April 2020, https://www.peoplemanagement.co.uk/long-reads/articles/is-ai-enemy-diversity.

Galloway, K 2017, 'Big Data: A case study of disruption and government power', *Alternative Law Journal*, vol. 42, no. 2, pp. 89–95.

Gaucher, D, Friesen, J and Kay, A 2011, 'Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality', *Journal of Personality and Social Psychology*, vol. 101, no. 1, pp. 109–128.

González, MJ, Cortina, C and Rodríguez, J 2019, 'The Role of Gender Stereotypes in Hiring: A Field Experiment', *European Sociological Review*, vol. 35, no. 2, pp. 187–204.

Hays, S 1996, *The Cultural Contradictions of Motherhood*, Yale University Press, New Haven.

Hegarty, P and Buechel, C 2006, 'Andocentric Reporting of Gender Differences in APA Journals: 1965-–2004', *Review of General Psychology*, vol. 10, no. 4, pp. 377–389.

Jiang, Y and Jones, S 2018, 'Corporate Distress Prediction in China: A Machine Learning Approach', *Accounting & Finance*, vol. 58, pp. 1063—1105.

Judge, TA and Cable, DM 2004, 'The effect of physical height on workplace success and income: preliminary test of a theoretical model', *Journal of Applied Psychology*, vo. 89, no. 3, pp. 428.

Kim, PT 2019, 'Big Data and Artificial Intelligence: New Challenges for Workplace Equality', *University of Louisville Law Review*, vol. 57, pp. 313–328.

Kulkarni, S and Che, X 2017, 'Intelligent Software Tools for Recruiting', *Journal of International Technology & Information Management*, pp. 1–16.

Lambrecht, A and Tucker, C 2020, 'Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads', *Management Science*, pp. 2966–2981.

Lewicki, RJ, Saunders, DM and Barry, B 2016, *Essentials of Negotiation*, 6th Edition, McGraw-Hill Irwin.

Lindsay, PH & Norman, DA 2013, *Human information processing: An introduction to psychology*, Academic Press.

Lim, KH, Benbasat, I and Ward, LM 2000, 'The role of multimedia in changing first impression bias', *Information Systems Research*, vol. 11, no. 2, pp. 115–136.

Mayew, WJ and Venkatachalam, M 2012, 'The Power of Voice: Managerial Affective States and Future Firm Performance', *The Journal of Finance*.

Miller, T 2019, 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, no. 267, pp. 1—38.

Mitchel, TM 1980, 'The need for biases in learning generalizations', Technical Report. CBM-TR-117, Rutgers University, New Brunswick, NJ.

Mittelstadt, BD, Allo, P, Taddeo, M, Wachter, S and Floridi, L 2016, 'The ethics of algorithms: mapping the debate', *Big Data & Society*, DOI: 10.1177/2053951716679579.

O'Neil, C 2016, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, United Kingdom: Penguin Random House.

Oreopoulos, P 2011, 'Why do skilled immigrants struggle in the labour market? A field experiment with thirteen thousand resumés', *American Economic Journal*, vol. 3, no. 4, pp. 148–171.

Preuss, A 2017, 'Airline pilots: the model for intelligent recruiting?', *Recruiter*, pp. 12–13.

PricewaterhourseCoopers (PwC) 2015, *Data driven: What students need to succeed in a rapidly changing business world*. Available at: <https://aechile.cl/wp-content/uploads/2015/02/PwC-Data-driven-paper-Feb2015.pdf>.

Purda, LD and Skilicorn, D 2015, 'Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection', *Contemporary Accounting Research*.

Raub, M 2018, 'Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices', *Arkansas Law Review*, vol. 71, no. 2, pp. 529–570.

Rudin, C 2019, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, vol. 1, pp. 206—215.

Russo, N 1976, 'The motherhood mandate', *Journal of Social Issues*, no. 32, pp. 143–153.

Sczesny, S, Formanowicz, M and Moser, F 2016, 'Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination?', *Frontiers in Psychology*, vol. 7, no. 25, pp. 1–11.

Tsai, CF and Chen, ML 2010, 'Credit Rating by Hybrid Machine Learning Techniques', *Applied Soft Computing,* vol. 10, no. 2, pp. 374—380.

Tsamados, A, Aggarwal, N, Cowls, J, Morley, J, Roberts, H, Taddeo, M and Floridi, L 2021, 'The ethics of algorithms: key problems and solutions', *AI & Society*, DOI: 10.1007/s00146-021-01154-8.

Van der Voort, HG, Klievink, AJ, Arnaboldi, M, and Meijer, AJ 2019, 'Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making?'*, Government Information Quarterly*, no. 36, pp. 27–38.

Van Esch, P, Black, S and Ferolie, J 2019, 'Marketing AI recruitment: The next phase in job application and selection'*, Computers in Human Behaviour*, no. 90, pp. 215–222.

Workplace Gender Equality Agency 2019, 'Gender Segregation in Australia's Workforce', *Australian Government WGEA*, <https://www.wgea.gov.au/data/fact-sheets/gender-segregation-in-australias-workforce>.

Zuboff, S 2019, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile Books Ltd, London.