Digit Working Paper No. 3
March 2022

digital futures at work research centre

Rapid Recruitment in Retail:

Leveraging AI in the hiring of hourly paid frontline associates during the Covid-19 Pandemic

William Hunt and Jacqueline O'Reilly

Digital Futures at Work Research Centre, University of Sussex Business School

Economic and Social Research Council

As part of the Digital Futures at Work Research Centre (Digit), this work was supported by the Economic and Social Research Council [grant number ES/S012532/1]

US UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

UNIVERSITY OF LEEDS
Leeds University Business School

The **Digit Working Papers** series is an open access resource of peer-reviewed papers from the Digital Futures at Work Research Centre (Digit). This series presents concept papers, findings and theoretical investigations around the digital futures of work that are the result of, and contribute to, Digit research. The working papers are intended to meet our core objectives of:

a. generating new knowledge, ideas and methods
b. engaging policymakers and practitioners through communications, knowledge exchange and impact
c. strong career development for researchers, and
d. ensuring long-term sustainability of the produced knowledge and the Centre.

Suggested citation for this report:
Hunt, W. and J. O'Reilly (2022) Rapid Recruitment in Retail: Leveraging AI in the hiring of hourly paid frontline associates during the Covid-19 Pandemic. *Digit Working Papers No. 03*, University of Sussex, Falmer. https://dx.doi.org/10.20919/ALNB9606

DOI: 10.20919/ALNB9606
ISSN 2755 – 1350
Publisher: Digital Futures at Work Research Centre (Digit)
Address: University of Sussex, Falmer, BN1 9SL
Editors:
Prof Jacqueline O'Reilly, University of Sussex
Prof Mark Stuart, University of Leeds

Views expressed in this working paper are those of the author(s) and not those of Digit.

About the authors at the time of publishing:
Wil Hunt is a Research Fellow at the Digital Futures at Work Research Centre (Digit). His research investigates how new digital technologies and AI are changing the world of work. https://digit-research.org/researcher/dr-wil-hunt/
Jacqueline O'Reilly is Professor of Comparative Human Resource Management at the University of Sussex Business School and Co-Director of the Digital Futures at Work Research Centre. https://digit-research.org/researcher/prof-jacqueline-oreilly/

Acknowledgements:
The authors would like to thank Julie Murphy (Executive Vice President of People, Walmart US) and all interviewees who took part in the research from Walmart US. The authors would also like to thank Lara Shock and Thea Powers at Walmart US for their help supporting the research. As part of the Digital Futures at Work Research Centre (Digit), this work was supported by the UK Economic and Social Research Council [grant number ES/S012532/1], which is gratefully acknowledged.

Abstract

Increased demand due to the Coronavirus pandemic created the need for Walmart to onboard tens of thousands of workers in a short period. This acted as a catalyst for Walmart to bring forward existing plans to update the hiring system for store-level hourly paid associates in its US stores. The Rapid Recruitment project sought to make hiring safer, faster, fairer and more effective by removing in-person interviews and leveraging machine learning and predictive analytics. This working paper reports on a case study of the Rapid Recruitment project involving semi-structured qualitative interviews with members of the project team and hiring staff at five US stores. The research finds that while implementation of the changes had been successful and the changes were largely valued by hiring staff, lack of awareness and confidence in some changes threatened to undermine some of the objectives of the changes. Reservations about the pre-employment assessment and the algorithm's ability to predict quality hires led some users reviewing more applications than perhaps necessary and potentially undermining prediction of 90-day turnover. Concerns about the ability to assess candidates over the phone meant that some users had reverted to in-person interviews, raising the risk of Covid transmission and potentially undermining the objective of removing the influence of human bias linked to appearance and other factors unrelated to performance. The impact of awareness and confidence in the changes to the hiring system are discussed in relation to the project objectives.

Key words:
Artificial intelligence/machine learning, predictive analytics, automated hiring systems, recruitment, technology adoption

Key Points

1. Walmart had implemented changes to the hiring system for store-level hourly paid associates during the Coronavirus crisis. Machine learning aimed to make hiring more data-driven and the removal of in-person interviews aimed to reduce risk and speed up the process. The changes also aimed to improve outcomes and remove human bias.

2. Lack of confidence in the changes and awareness about the technology used threatened to undermine the objectives of the changes in some cases. However, while awareness of the technology used might increase buy-in from some users, others may remain sceptical unless they see improvements.

3. Lack of awareness and confidence in the algorithm used to rank candidates meant some users did not take full advantage of the technology used and led some to review more applications than perhaps necessary: undermining the objective of speeding up hiring and making it more data-driven.

4. Lack of confidence in the algorithm and pre-employment assessment and hirers' ability to assess the quality of candidates over the phone, led some managers to revert to in-person interviews: potentially increasing risk of Covid transmission during the pandemic and undermining efforts to reduce the influence of human bias related to appearance.

5. Misalignment between the qualities hirers and the algorithm are trying to predict may explain some users' reservations about the systems capabilities. Increasing awareness about the technology used and the qualities the system aims to predict may go some way to building trust in the system.

Contents

Tables and figures

Abbreviations

ADM	Algorithmic Decision Making
AI	Artificial Intelligence
AHS	Automated Hiring Systems
API	Application Programming Interface
EDI	Equality, Diversity and Inclusion
HH	Hiring Helper
ML	Machine Learning
OGP	Online Grocery Pick-up
UI	User Interface

1. Introduction

The focus of this working paper is the Rapid Recruitment project implemented for store-level hourly paid associates at Walmart, USA. The project sought to leverage existing artificial intelligence (AI) and machine learning (ML) technology to speed up the hiring system for hourly paid associates and make hiring decisions fairer and more data-driven. Drawing upon a small research project consisting of qualitative interviews with key personnel participating in the development of the project and core users of the hiring system, this working paper outlines user responses to changes to the hiring system and the implications of these for the organizational objectives of the project.

The research comprised of qualitative interviews with 14 Walmart employees with a range of roles in relation to hiring at the organization, including: seven employees from home office with a range of responsibilities for the development and implementation of the project; five people leads and store managers, and two recently recruited hourly paid associates. The research involved in-depth semi-structured video interviews with respondents (n=14) lasting from 30-60 minutes, supplemented by biweekly catch-ups with the business sponsor for the project and follow-up emails with respondents for fact checking information. As the research was conducted during the Covid pandemic 2020, all interviews were conducted on Zoom.

The paper starts by summarizing key literature in the area before providing an overview of the project and its main objectives. We then detail user response to changes in the hiring system in relation to the three broad objectives of the Rapid Recruitment project. Based on this analysis we conclude that in order to ensure changes to the hiring system are successful, fair and unbiased it will be important to:

1) Build trust in the Hiring Helper by increasing awareness and ensuring staffing recommendations are reliable;
2) Continue to monitor and assess the algorithm and its output to ensure that there is no bias in the system;
3) Continue efforts to identify and address potential systematic sources of bias in other parts of the system (e.g. that telephone interviews do not systematically favour some groups of candidates over others)

Clearly given the volume of staff being recruited in retail these types of automated systems will become more common; however, to ensure their effectiveness requires constant monitoring of the performance and outcome of automated decisions and the way staff interact with these new systems.

2. Algorithmic decision making and automated hiring systems

Advancements in computer processing power, data science and artificial intelligence (AI) have greatly expanded the range of tasks that computers can be put to (Brynjolfsson, Rock and Syverson, 2017). Algorithmic decision making (ADM) – whereby decisions and predictions can be automated based on predefined rules and goals with little or no human involvement (Allen and Masters, 2020; Lindebaum, Vesa and Den Hond, 2020) – has been applied to advertising, policing, criminal sentencing and, increasingly, hiring (Lambrecht and Tucker, 2019; Eubanks, 2018; O'Neil, 2016; Bogen and Reike, 2018; Chen, Ma, Hannák and Wilson, 2018).

With ADM, decisions or predictions can be made very quickly, often in real time, greatly expanding the possibilities to upscale decision-making processes. Proponents argue that computers using AI can perform tasks and make decisions quicker more effectively than humans and are better able to detect patterns in data (Upadhyay and Khandelwal, 2018; Van Esch and Black, 2019; Newman, Fast and Harmon, 2020; Frey and Osborne, 2017). Furthermore, ADM can potentially help bypass known human biases that lead to exclusion and discrimination such as: validity illusion, affinity bias, status quo bias, noise shocks, and interpersonal bias (Bogen and Reike, 2018; Howser, 2019; Cowgill, 2018).

2.1 Automated hiring systems

Automated hiring systems (AHSs) using AI and machine learning (ML) are increasingly being applied to all four stages of the recruitment and hiring pipeline: advertising/sourcing, screening, interviewing/assessment, and selection and onboarding (Raghavan, Barocas, Kleinberg and Levy, 2020; Brione, 2020). Real world examples of applications of AI and ML identified in the literature (e.g. Bogen and Reike, 2018; Albert, 2019; Heric, 2018; Ernst, Merola and Samaan, 2018) include:

- Head hunting tools, such as those offered by Entelo and LinkedIn that proactively search a wide range of data sources to surface potential candidates, particularly for higher-skill positions, and attempt to predict company fit, diversity and likelihood of moving jobs. High profile users include American Express, Atlassian and Oracle;
- Job description optimization software, such as Textio, that use natural language processing to analyse text in order to predict expected candidate pool and identify language that may alienate protected groups. Reported adopters include P&G, Johnson&Johnson and Twitter;
- Algorithmically controlled targeted advertising of vacancies using platforms such as Facebook, Google and LinkedIn;
- Vacancy-candidate matching tools, such as those used by LinkedIn and Monster.com. ZipRecruiter uses 'content-based' (based on user click and interests) and 'collaborative' (based on the interests of similar people) filtering and employing a 'recommender' style system like those used by Netflix or Amazon;

- CV screening software, such as Ideal, which can instantly review and screen a large number of CVs or applications and rank the 'best' candidates. Amazon, IBM and Goldman Sachs are adopters of such technology;
- AI-powered psychometric testing – Koru and Pyemetrics are high-profile examples that use games and psychometric-based assessments in order to predict cognitive, behavioural and personality traits expected to be good predictors of fit and performance. Unilever, PwC and Accenture are high-profile users;
- Video screening – For example, HireVue uses video interviews to measure human traits from non-verbal cues, vocal cues and language use, and employ ML to predict workplace performance. Vodafone and Intel are two high-profile users of video screening software;
- Candidate engagement chatbots using natural language processing can be used to engage applicants and provide quick responses to questions and even provide instant feedback to applications. H&M and eBay are reported to be users of this type of technology;
- Automated scheduling software uses AI to pick up on scheduling expressions and automatically executes admin tasks, allowing recruiters to focus on more essential tasks.

In many cases, employers use a combination of tools and technologies, both in-house and/or from third party vendors (Bogen and Reike, 2019). Unilever, partnering with Pyemetrics, employed a two-stage process involving a game followed by a video interview to screen candidates (Heric, 2018; Marr, 2018; Booth, 2019). ML was used to predict career success using traits measured from facial expressions, body language and word choice. Similarly, German company SAP developed and tested a screening process using CV screening and two algorithmically assessed pass/fail screening assessments – a ten minute cultural fit test and a 20 minute situational judgement test – and a day long assessment (Hopping, 2015).

2.2 Automated hiring systems, algorithmic decision-making and bias

Such systems have the potential to ensure hiring is fact based and driven by data, not intuition (Van Esch and Black, 2019). This can potentially improve the quality and diversity of hires. Diversity in the workforce is not only an ethical goal, but hiring diverse employees stimulates innovation and creativity, helping to improve productivity and competitiveness (Hunt, Layton and Prince, 2015; Pessach et al., 2020). Thus, AHSs can potentially have wide-ranging organizational performance benefits. However, while commercially available AHSs commonly claim to circumvent human subjectivity by ranking individuals on 'objective' measures, these claims are rarely scrutinized by academic research (Raghavan et al., 2020; Sánchez-Mondero, Dencik, and Edwards, 2020).

Conversely, some have argued that ADM has the potential to reproduce and even amplify human biases, while giving only the illusion of objectivity (Eubanks, 2018; O'Neil, 2016; Newman et al., 2020; D'Alessandro, O'Neil and La Gatta, 2017). Mehrabi et al. (2019) identify no fewer than 23 potential sources of bias in ML and outline four broad types of discrimination in data: direct, indirect, systemic and statistical. To which, a further four can be identified in the Data Science literature (Kamishima, Akaho, Asoh and Sakuma, 2012; Turner Lee, 2018). These can be synthesized as follows:

- **Direct prejudice/discrimination** – use of a sensitive characteristic in the prediction model;
- **Indirect prejudice/discrimination** – the prediction model includes a characteristic that is highly correlated with a sensitive characteristic;
- **Latent prejudice/discrimination** – the prediction model includes a characteristic, or characteristics, that are partially correlated with a sensitive characteristic;
- **Systemic/structural discrimination or bias** – where the culture or policies of an organization lead to discrimination of certain groups (e.g. hiring people with similar experiences or interests) or bias that stems from existing patterns of disadvantage (e.g. using *alma mata* to rank candidates when access to different universities is known to be unfair);
- **Statistical discrimination** – group averages used in data models (e.g. because a measure of interest is hard to observe) leading to unfair treatment;
- **Underestimation** – where the training dataset is too small for the ML to learn an accurate prediction model. While any learned bias is unintentional, unfair decisions may result when applied to the wider population;
- **Negative Latency** – unfair sampling or labelling in the training data. For example, if a given organization's historical recruitment decisions have been biased, the selection bias in the training dataset can lead the ML to learn this bias. This issue is difficult to overcome because it is not known how many of the past rejected candidates would have made good employees.

While the potential advantages of AHSs and the use of ML in recruitment are real – Unilever reportedly saved 100,000 recruiter hours using a two-stage AI-based candidate screening system and SAP reportedly saved a projected £250,000 in the first year following implementation (Heric, 2018; Booth, 2019; Marr, 2018; Hopping, 2015) – some high profile examples underline the potential issues. For example, an algorithm used to target STEM job ads on Facebook was shown to target men almost exclusively because cost-effectiveness optimization effectively screened out young women who command a higher advertising premium (Lambrecht and Tucker, 2019). In 2017, Amazon famously abandoned development of a CV screening algorithm because it screened out CVs that used the word "women's" or where applicants had attended all female colleges because of bias in the training data where the company had historically hired men for developer and technical posts (Brione, 2020; Dastin, 2018). However, while fairness and reducing bias in ML and ADM are established topics in the data science literature, understanding how use of these technologies play out in hiring practices is relatively understudied. And, while the sheer number of AHS providers is testament to the extent to which the technology is starting to be applied to hiring, there is little academic research as to how such technologies are being applied in practice and the implications for organizational objectives.

3. The Rapid Recruitment project

3.1 Background and objectives

Staff turnover in the retail industry

Staff turnover in the retail industry is high compared to other industries: the total annual turnover in retail averaged at 58.4% in 2019 compared to a 45.1% average across all industries in the US.[1] Similar patterns are also found in the UK with employee turnover rates second only to the hospitality sector and arts, entertainment and recreation.[2] Walmart is not immune to this with respondents from the Home Office and from stores noting that staff turnover among store-level hourly associates was a significant challenge, particularly among new hires. In an industry where customer service is key, time and resources spent recruiting and training staff can be substantial and turnover can affect quality of customer service.

Fairness and diversity

Ensuring fairness and diversity are not only ethical goals, they can also help improve productivity and competitiveness by enhancing innovation and creativity (Hunt et al., 2015; Pessach et al., 2020; Cowgill, 2018). This view was reflected in the views of interviewees at Walmart who cited equality, diversity and inclusion (EDI) as important for achieving organizational goals:

> *We've got an overarching vision of everyone included, we believe that when associates are and feel welcomed, comfortable and safe, they're engaged and empowered to be high performers, to deliver better service for our customers and our members…We believe that diverse teams outperform non-diverse teams, but only when they're in an inclusive environment.*
>
> *(EDI Champion)*

> *We're a company that's been focused on diversity and inclusion as long as I've been in management with the company and well before that… I think we've always tried to recruit diverse candidates. I think we've always tried to promote people based off of their ability, not so much, you know, fitting into a certain mould.*
>
> *(Joe, Experienced store manager at a supercentre)*

The importance of EDI to Walmart is also evident in a number of wider organizational schemes and policies aimed at supporting EDI, including the setting up of a Global Office of Culture, Diversity and Inclusion and schemes such as the Center for Racial Equity Award grants.[3] Thus, the ability to identify and retain quality personnel in a speedy and fair way is an important organizational objective.

[1] https://www.bls.gov/news.release/jolts.t16.htm

[2] https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/adhocs/10685employeeturnoverlevelsandratesbyindustrysectionukjanuary2017todecember2018

[3] https://corporate.walmart.com/newsroom/2021/02/01/the-walmart-org-center-for-racial-equity-awards-over-14-million-in-first-round-of-grants

The Coronavirus Catalyst

Walmart employs 1.5 million workers across their 4,100 US stores. The advent of the Coronavirus crisis in early 2020, and the ensuing surge in consumer demand, created the need to onboard tens of thousands of workers in a short period. The organization recruited 460,000 new associates in the period from March to November 2020. This amplified the need for a fast, effective and fair hiring system for store-level associates. In response, Walmart brought forward existing plans to update the hiring and onboarding system that had been under development for two years.

The overarching aim of the Rapid Recruitment project was to leverage AI-enabled technology to improve retention and quality of new hires among frontline hourly paid associates. This was to be achieved through speeding up the hiring process and making hiring decisions less subjective and more data-driven.

Objectives

The reported objectives of the project were to:

1) Make 'better' hiring decisions – in terms of length of tenure and employee performance;
2) Minimize the influence of human bias in hiring decisions;
3) Speed up the hiring and onboarding system to enable the recruitment of a greater number of associates within a shorter time period (aim 48 hours);
4) Protect the health and safety of hiring staff and applicants during the pandemic by reducing the need for face-to-face contacts.

3.2 Changes implemented

These objectives were to be achieved by the implementation of two main changes to existing hiring and onboarding processes for hourly paid store-level associates:

1) Organizational changes to the hiring process (detailed below); and
2) Changes to background checks and onboarding processes aimed at reducing the amount of time between job offer and start date in store.

While this report focuses on the former, the latter involved a more streamlined process for background checks, enabling new hires to start after part 1 of the background check had completed and enabling new hires to complete some onboarding activity remotely before their first day in store. This reduced the chances of new hires finding alternative jobs before background checks were completed.

Changes implemented to the hiring processes included:

1) Development of an algorithm using machine learning to help sort applicants in the Hiring Helper[4] hiring system used by people leads and store managers;
2) Organizational changes to the hiring process and hiring management system, including:
    a. Replacing in-person interviews with shorter telephone interviews;

[4] The Hiring Helper is the hiring database system used by people leads and store managers to manage job requisitions and applications.

b. Changes to the user interface (UI) in the Hiring Helper;
c. Changes to the guidance around the use of the Hiring Helper to encourage users to make better use of the sorting technology; and
d. Enabling people leads and managers to make over the phone job offers.

Together these changes aimed to make hiring decisions more 'data-driven': based solely on information that was genuinely related to these desired outcomes and not influenced by other, less relevant factors.

By making hiring decisions more data-driven the team hoped to both improve hiring outcomes and minimize the influence of human bias. These two objectives were often seen as being two sides of the same coin. The changes to the Hiring Helper UI were aimed at making candidates near the top of the list and the reasons for their position more prominent encouraging trust and confidence in the algorithm's recommendations, and enabling hiring managers to make job offers over the phone meant that hiring could be performed much more quickly and with no need for in-person contact.

3.3 Development of the algorithm and the Hiring Helper

The project team sought to leverage existing work on the algorithm that had been in development since 2018 and had been incorporated into the Hiring Helper in 2019. The project was described as multi-year journey aimed at improving the existing hiring and staff management system.

The objective was to make hiring decisions more 'data-driven' and less biased, by making recommendations based solely on information that could be shown to predict organizational outcomes and not influenced by other, less relevant factors. In other words, the aim was to help people leads and store managers to identify the 'best' candidates in terms of potential performance and the likelihood of staying in post.
Machine learning technology was used to train the algorithm to predict which candidates would make a good hire and sort them in the list of candidates presented to people leads when opening a new job requisition. A number of different target variables ('labels') related to positive hiring outcomes were considered, including different measures of employee performance and length of tenure.

Predicted 90-day turnover (i.e. likelihood of staying in post for longer than 90-days) was chosen as the target variable because it is an objective measure, available for all hires. As such a business case could be made for the value of improving 90-day turnover in future hiring decisions. Other performance measures, such as information from performance reviews, were either subjective or not available for all staff.

Fair Aware Classifier

The algorithm was trained on historical data for an existing hire cohort (500,000-600,000 hires) using a Fair Aware Classifier,[5] using more than 30 data points. Measures used in the algorithm included metrics related to applicants' employment history, their scores on the pre-employment assessment and geographical measures (e.g. distance from store). The Hiring Helper also applied business rules based on a number of variables including job preference, shift availability, if applicants had failed the pre-employment assessment, if they had previously had employment at Walmart terminated and if they were ineligible for the particular requisition (e.g. if they were a minor). The algorithm did not make use of protected characteristics, such as age, gender or race, for legal and ethical reasons.

A Fair Aware Classifier approach was chosen because it helped ensure that any variance in measures used in the algorithm that was correlated to protected variables would be 'stripped out' of the data while retaining variance that helped predict the target outcome.

The large amount of historical data available for the organization was seen as a major advantage in developing the algorithm, as the greater the amount of data that is available the more accurate the resulting algorithm. Once the algorithm had been trained on the training dataset it was piloted in two markets with 25 stores before being rolled out for general use in stores. Prior to this, the list of potential candidates presented for any given requisition could be sorted by people leads on any number of variables, such as name, distance from store or date of application.

Building trust and confidence

Building trust and confidence in the algorithm's ability to predict good hires was identified as an important factor in the successful implementation of the system. In order to achieve this, it was seen as essential to ensure that output was: 1) accurate, 2) trustworthy, and 3) transparent. The project team described building trust as an ongoing journey involving listening to user concerns and the ongoing development of the algorithm and UI.

> *We don't expect them out of the gates to just follow our direction. Right? We're doing this and we're on a journey to learn, learn and refine, learn and refine, until we get it right. And we'll never get it perfect but we should be able to get it better than they [managers] would, at some point because we're leveraging all these data elements.*
>
> *(Project Lead)*

In addition, it was also necessary to make the hiring and talent management system 'portable' using Application Programming Interface (API) so that future adjustments and changes to the hiring and talent system can be made more easily.

[5] A fairness aware classifier is a type of machine learning algorithm that aims to accurately predict a target variable while reducing the influence of indirect prejudice, which can occur when a sensitive characteristic is correlated with a variable used in the model even if the sensitive characteristic itself is not used in the model (Kamishima et al., 2012).

3.4 Organizational changes to hiring processes and changes to the hiring management system

Prior to the implementation of the project, candidates for hourly paid roles would have been interviewed in-person on site. The onset of the coronavirus crisis led to the team rethinking and bringing forward existing plans to make organizational changes to the hiring process. The in-person interview was removed and replaced with a shorter, less detailed, telephone interview.

Additional guidance was disseminated via Amp[6]: 1) advising people leads and hiring managers to use the list presented in the Hiring Helper as a starting point for identifying suitable candidates, and 2) enabling them to make job offers over the phone if they felt a candidate was suitable based on their application and information gathered over the phone.

Changes were also made to the Hiring Helper UI in order to make candidates ranked higher in the list, and the reasons for their position, more prominent. Whereas candidates in the previous UI were listed in tabulated form with candidates as rows and key metrics as columns, the updated UI displayed the top five candidates as tiles on a carousel displayed horizontally from left to right, with users needing to scroll along the tiles to see candidates further down the list. An 'explain to me' button was also added to the new UI that listed the main pros and cons that explained the candidates' position in the list. These changes to the UI were intended to increase trust in the system and make it more user friendly.

It was hoped that increasing trust in the Hiring Helper would help achieve objectives 1 and 2 (p13, this working paper) by making hiring decisions more 'data-driven' and less prone to conscious or unconscious human biases related to the appearance of candidates. If users have trust in the Hiring Helper, they can call up candidates sorted at the top of the list by the algorithm and offer them the job on the spot if they are suitable. This also helps speed up the hiring process and minimize transmission of Covid-19 by reducing the need for the scheduling of multiple in-person interviews. At the time of the interviews, while the algorithm was in use in all stores the updated UI had not yet been rolled out to all stores.

All of these changes were intended to make hiring choices more data-driven, by removing opportunities for unconscious bias and elevating the role of the algorithm in guiding hiring choices. This way decisions about who to interview would be guided by predictive analytics, based on metrics related to turnover, and information from candidates' application forms. Thus, job offers would be less likely to be influenced by factors unrelated to turnover, such as human bias related to appearance, personal characteristics and pre-conceived ideas about candidates' capabilities.

[6] The organization's internal messaging system.

4. Outcomes and user response

4.1 Overview of user response

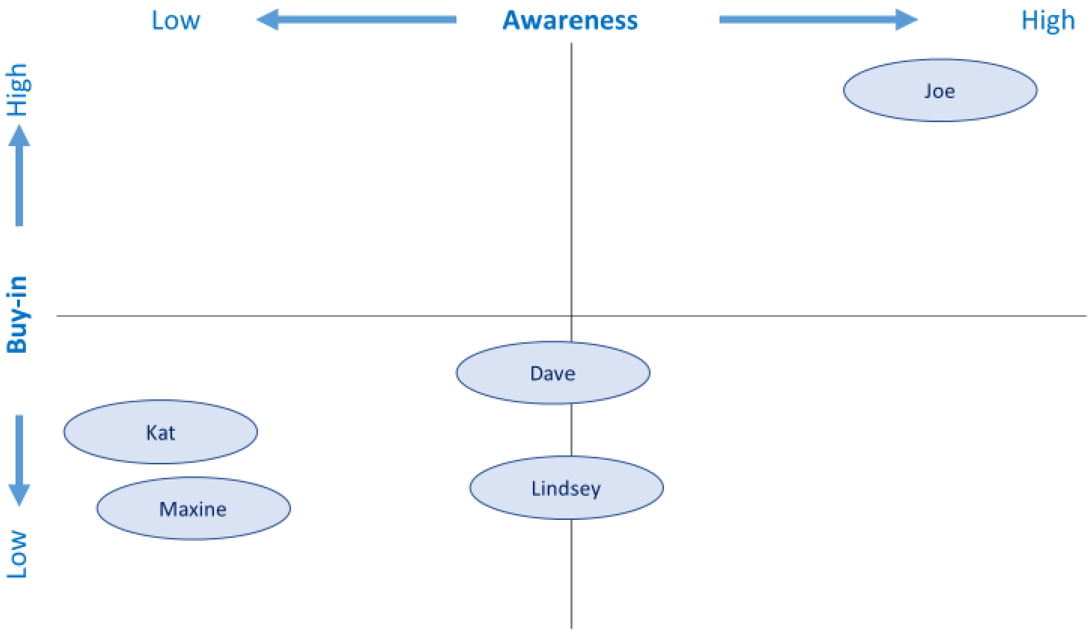
The people leads and store manager interviewed generally accepted the recent changes made to the system around the start of the coronavirus crisis. All accepted that the removal of the in-person interview was a necessary step to protect staff and applicants from transmission of the virus and all appreciated the timesaving advantages.

However, awareness of the technology used in the system and confidence and trust in its capabilities varied. Behavioural responses to the changes made to the system and use of technology can be thought of varying along two continuums:

1) Extent to which they were aware of the technology involved and its objectives;
2) Level of buy-in to the new system and confidence in the Hiring Helper's ability to make good hiring recommendations.

Figure 1 shows where individual responses fit in this two-dimensional space and Table 1 summarizes the responses of individual users. While from the figure it might appear that buy-in is correlated with awareness, it is clear from user's reports that buy-in and trust in the system has more to do with the perceived ability of the system to identify good hires (compared to the old system) than awareness of the technology involved. While greater awareness of the technology involved in the system could increase buy-in from some users, other users may remain sceptical. These issues are discussed in more detail in the following sections where user responses are discussed in relation to the main objectives of the project.

Figure 1 – Behavioural response to the Hiring Helper and Rapid Recruitment project



Notes: user response is plotted along two dimensions from low to high in terms of (i) awareness of the new system and the technology involved and (ii) the extent to which they trust the system and use it as intended

Table 1 – Summary of individual user responses

Respondent	Behavioural response
Joe	**Experienced store manager in a mid-size supercentre in an urban market** Aware that algorithm is attempting to predict quality and longevity of applicants; Generally feels that well trained manager can make better decisions but had become convinced that the HH was making as good, if not better, recommendations; felt that this was particularly useful for new managers and to achieve much quicker hiring times; used recommendations to focus attention to applicants at the top of the list; has a view that, despite best intentions, human decisions are subject to biases and ADM can help minimize that.
Lindsey	**People lead in a high-volume supercentre in a large rural market with more than 7 years' experience in an HR role** Partially aware of the technology used in the HH; Understood that it was attempting to predict good hires but was sceptical of its ability to do so; gave examples of some at the top of the list who turned out not to be good hires and some appearing far down the list that they felt would have made good associates; believed that the main aim of removing the in-person interview was to avoid Covid transmission; felt that this had the potential to remove some human bias but felt that some of the advantages of in-person interviews, such as judging honesty through body language, were lost and that appearance was a good indicator of professionalism; had reverted to in-person interviews as had been 'burned' by some decisions.
Kat	**People lead in a complex supercentre in an urban market, only in post for 1 year** Tends to go through all the applicants on the first and second page on the list (NB – store is on the old UI) focusing on those that have appropriate availability and who scored good or excellent on the assessment; mostly goes through the list in order but reviews up to 20 before deciding who to call for a telephone interview. Out of those called, applicants finally hired could appear anywhere in the list. Reported sometimes starting at the bottom of the list (of the 5-20) when they need to fill a number of requisitions because names at top might appear in several requisitions (but noted that sometimes these had applied a long time before). Assumed position on the list was related to when the applicant applied (most to least recent). Believed that the removal of in-person interviews was for safety reasons related to Covid and to speed up the process but that there were no other reasons. Felt that the pre-employment assessment was a good predictor of turnover. Felt that you can get a good sense of attitude and customer service skills over the phone and that this was likely to make hiring fairer because you cannot see the applicant (although recognised this might not suit every applicant).
Dave	**People lead at a supercentre in a large rural market, more than 9 years of HR experience** Likes to look at all the applications in the requisition. The store uses the 'explain to me' function 30% of the time. Reported that some managers were sceptical of the recommendations and like to work things out for themselves. Understands that the list is sorted by an algorithm and that it takes different metrics into account, but unsure of the metrics and technology used. Felt that the recommendations were not always accurate, gave examples of recruiting applicants rated poor who turned out to be great associates; concerns were partly about the assessment, which he felt was subjective and might not be good for those who are nervous or with poor digital skills. While he noted the speed and safety benefits of removing the in-person interview, he felt that the current process has lost some of the 'human element'.
Maxine	**Experienced people lead at a supercentre in an urban market** Was unaware that the HH used AI technology but understood that the system took into account a number of different factors and business rules. Reported going through the list as requested, but was sceptical of the recommendations it made. Wondered to what extent number one on the list was better than number two, or even number 20, and gave examples of people further down the list that they felt made good associates. Her main feeling was that the assessment was not necessarily accurate in all areas.

4.2 Speeding up hiring and making it safer during the pandemic

Users generally understood and had adopted the new hiring process, although some aspects had been adopted to varying degrees. All had stopped carrying out in-person interviews and were instead carrying out shorter telephone interviews. However, one respondent reported that some hiring managers had recently reverted to in-person interviews as they lacked confidence in the ability to identify good hires in the absence of face-to-face interaction with candidates.

All of the people leads and the store manager interviewed understood that changes to the system were intended to speed up the process and minimize the transmission of Covid-19. In fact, at least one respondent was under the impression that these two objectives were the only reasons for the changes to the hiring system.

> *The objective was to keep us safe I would say. You have safety precautions for everyone, so you didn't have as many face to face kind of contacts with people.*
>
> *(Kat, People lead)*

All interviewees reported that the changes to the hiring system, along with the changes to background checks and onboarding processes, had indeed sped up the hiring process from one to two weeks to as little as 48 hours. This was universally valued by the hirers interviewed and all felt that removing the in-person interview was a necessary step to protect staff and candidates during the pandemic.

> *Well, the advantages are it's safer as far as everyone, you know, possibly, for health issues. … And I would say also, the advantages are being able to quickly get candidates in. This process is definitely a lot quicker. So I will say that I do love that about it because you don't have to wait for a person to come in and their availability and interview. You can do everything the same day. So it makes the process a lot easier to fill positions.*
>
> *(Kat, People lead)*

While all respondents felt that carrying out interviews by telephone saved time and may help eliminate unconscious bias related to appearance, others had reservations about the recommendations of the Hiring Helper and the ability to assess candidates over the phone. These reservations – discussed in the following section – led some to review more applications than necessary or to revert to in-person interviews. For example, one respondent (at a store that was still using the old Hiring Helper UI) was unaware of the technology that went into the sorting of candidates, or that those further up the list were predicted to be good hires. They continued to go through 10-20 candidates in the list in much the same way as previously, sometimes starting at the bottom of the list.

> *So the first thing I would do is go through the profile open up a requisition. I would then go through all the applicants on the first to second page. Those are usually the ones who have just applied. Because most of the time they have not found the job yet and they're still, you know, currently looking. And I*

> *go through and check their availability. So I'll select the individuals off of their availability first. Okay, so once I get to about 20 people I've selected. I then go into the selections and do a deeper dive into the people specifically.*
>
> *(Kat, People lead)*

Another reported that some managers in the store had reverted to in-person interviews as they felt they had been 'burned' by one or two bad hires.

> *So the first person I come to that meets my criteria, and then I've looked at the more details and I like them, I just go ahead and call them. And I start working and kind of get a feel for him on the phone a little bit and then, if I'm unsure, I'll still call them in to come in for a face to face, even though there's not an official interview process. And we started doing that on a few just because we've been burned on a few.*
>
> *(Lindsey, People lead)*

Clearly, lack of awareness and/or confidence in the system had the potential to undermine the objectives of the changes, as users may spend longer going through the list of candidates than necessary, and in-person interviews meant more physical contacts increasing the risk of Covid transfer.

4.3 Making better hiring choices

Removing the in-person interview and elevating the role of the list in the shortlisting of candidates were intended to make hiring decisions more data-driven and less influenced by human bias, thus improving outcomes. Advocates argue that ML can identify patterns in data and predict outcomes far more effectively than humans (Upadhyay and Khandelwal, 2018; Van Esch and Black, 2019; Newman et al., 2020; Frey and Osborne, 2017). Walmart wanted to take advantage of this capability to improve 90-day turnover rates.

While all respondents had adopted the changes and some felt the changes had improved outcomes, reservations about aspects of the system threatened to undermine the objectives in some cases. Reservations included:

- Concerns about the way in which the algorithm incorporated some information (e.g. employment history) in the list;
- Reservations about the pre-employment assessment and its apparent prominence in the algorithm;
- Concerns about the ability to properly assess candidates over the phone and losing the 'human element'.

Awareness and trust in the changes to the hiring system

While users appeared to be unaware that making better hiring choices was one of the objectives of the recent changes, most recognized that the results of the pre-employment assessment attempted to predict the quality of candidates on some measure. Only one seemed to be aware that AI was used in the system, but several recognized that the system used an algorithm that took into account a number of factors, including assessment scores, shift preferences, distance from store and if they were ex-military.

> *I don't know in depth what the algorithm or whatever is and how it sorts the list. I just know how the list comes to me, and what order makes it more functional for me. But I've never had anybody just explain all the metrics that go into how the sorting is done. … But it's based on the assessment, the ones that score excellent are at the top and then it goes to good and then it goes down to poor.*
>
> *(Dave, People lead)*

Awareness of the fact that the list purportedly predicted 'better' hires did not necessarily mean that users were confident of its ability to do so. Two of those interviewed noted concerns about how the algorithm incorporated employment history into rankings, finding candidates with multiple short periods of employment (or 'job hoppers') high up the list. Some questioned how good the pre-employment assessment was at assessing the quality of candidates, particularly if candidates filled it in quickly or lacked digital skills. More than one respondent reported finding candidates near the top of the list who they felt would not make good associates or found candidates lower down the list who they were convinced would make good associates. Indeed, three of those interviewed felt that their reservations were proven to be correct by subsequent hires.

> *There has been times that, you know, they say, 'oh, try to hire somebody excellent or good or', you know, whatever. But I have, you know, based on interviews and people coming in, we've hired candidates that scored in the poor category that actually came in and are you know extremely, extremely, great associates. I can think of one that within three months of being here was a 'happy to help' associate for a month and, you know, hadn't even got their discount card yet.*
>
> *(Dave, People lead)*

In addition, one respondent felt that removing the in-person interview impeded their ability properly assess candidates:

> *You get a lot from that first impression when you see the person. And not being able to get that one-on-one interaction, like. There were a few that we hired, that sounded great on the phone and they came in we're like 'oh my gosh, who hired this person?', and then of course, you know, yeah, they were just terrible. Some of them.*
>
> *(Lindsey, People lead)*

They felt that appearance could be a useful indicator of a candidate's professionalism and that body language could help assess honesty and act as a cue for interviewers to probe further during interviews. Another respondent, while seeing some advantages to telephone interviews, lamented the loss of the 'human touch'.

> *I think when you put a human touch to the benefits, telling people in person, or having that personal touch to what the benefits and advantages of working at Walmart are, is way better than on a sheet of paper or in an ad. … It also makes them understand that our stores are tied to our community. We're a community store. We're Walmart, but we're a store of the community. And when you have that personal touch its 'Hey, I want to work there', 'I want to*

> *work there and have all these great things’ that ‘the people there are wonderful’. You know, that's the thing.*
>
> *(Dave, People Lead)*

However, not all users had reservations about the pre-employment assessment.

> *I check, first of all their assessments, I want to make sure that they scored good or excellent on their assessment because that definitely affects us, as a store, as a whole. It's known to have a higher turnover for people that do poor on the assessment. So I personally only pull from good or excellent.*
>
> *(Kat, People lead)*

The project team reported that analysis of the pilot scheme indicated that the algorithm was good at predicting whether associates were likely to leave in the first 90 days in post. However, they noted that this was difficult to confirm because turnover rates during the pandemic were unlike a normal year. This sentiment was echoed by some people leads and store managers who felt that removing the in-person interview had made hiring more data-driven and removed potential human bias related to appearance.

> *Like I said, it's been better for us as far as turnover. So personally, I would say this has been better. … Our numbers went down here… I think, maybe like to 30-40% possibly [from 70%], but it’s hard to tell [because of Covid].*
>
> *(Kat, People lead in a complex supercentre)*

One experienced manager felt that the algorithm was at least as good at identifying suitable candidates as most trained managers and felt that turnover had improved as a consequence. However, he noted that this was hard to verify due to the erratic effects of the pandemic. The manager showed a good level of awareness of the technology involved in the algorithm:

> *Well, I think it’s better. I honestly do. I think that. I think it's unbiased. And I think that's what makes it better. Because it has no preconceived ideas of anything about the individual themselves. It's just looking at the facts. So it's, it's giving you a better feel for their potential versus anything that you may bring to the table when you're doing that interview.*
>
> *(Joe, Store manager)*

And where there were reservations about the Hiring Helper’s ability make good predictions, these were mostly felt to be issues that could be corrected.

> *Yeah, and like I said, it's a lot better. I mean, it’s gotten better over the past, you know, few years. It's just we need a few more improvements. So, which I know that they're working on some of that.*
>
> *(Lindsey, People lead)*

> *I kind of think, ‘well, that predictive could get better’. You know, I think that assessment is, is kind of subjective and I think that shouldn't put somebody lower on the list.*
>
> *(Dave, People lead)*

Lack of awareness and/or confidence in the system affected the way in which users implemented the changes. While all had moved to telephone interviews, not all were using the list in the intended way with some interviewing more candidates than had been envisaged by the project team and one more or less ignoring the Hiring Helper altogether.

One respondent had a high level of awareness and confidence in the system used the system as intended: calling candidates at the top of the list (if they were eligible and shift preferences matched) and offering the job if happy with the candidate. Those with less confidence in the rankings, either continued to use the system as intended or preferred to reserve judgement until they had reviewed applications in more depth. One user was apparently unaware that the Hiring Helper attempted to predict good hires, believing the list to be based on recency of application, and continued to work through all candidates from the first two pages of list.

Thus, while users were less likely to be influenced by factors such as unconscious bias related to appearance or other characteristics, they were not making full use of the predictive analytics employed in the algorithm to help shortlist candidates. This potentially undermines the objective of making hiring decisions more data-driven, as some users were ignoring or sidestepping the recommendations of the Hiring Helper. To some extent the incongruence between the Hiring Helper's recommendations and users' own assessments of suitability could be rooted in a misalignment between how they interpret what makes a 'good' hire. Whilst the algorithm attempts to predict 90-day turnover, hiring managers and people leads may be assessing candidates on other qualities, such as potential performance, customer service, attitude or how they would 'fit' with the rest of the store team.

4.4 Making hiring fairer and less prone to bias

Making hiring fairer and less prone to human bias was another major objective of the Rapid Recruitment project. Bias, particularly human bias, was seen by the project team and by people leads as a both potential impediment to making better hires and to achieving EDI goals of the organization. By removing the in-person interview and making greater use of the algorithm, it was hoped that this would also make hiring less prone to preconceived ideas about a candidate's suitability based on appearance and other personal characteristics unrelated to performance.

Approach to fairness and bias

Fairness was often conceptualized in terms of giving everyone an equal chance and making sure the best person is selected for the job, rather than in terms of making sure that the profile of associates reflects the profile of applicants or the local population. This corresponds more to conceptualizations of 'individual fairness' rather than 'group fairness' (Bogen and Reike, 2018) and links to the legal concept of 'disparate treatment' rather than 'disparate outcomes'. This focus on individual fairness was voiced by the project lead:

> *How can we stop hiring people at the beginning of the alphabet? How can we give everybody a fair chance? Right, like let's assess actual skills, work*

> *history, things that are going to matter and be relevant to the performance and if someone's going to stay at Walmart.*
>
> *(Project lead)*

People leads and hiring managers, also often saw fairness as being about getting the best person for the job:

> *To give everybody an opportunity. No matter. … based on the tools and the process, to give everybody an opportunity if they qualify.*
>
> *(Maxine, People lead)*

Although most recognised that this also meant ensuring that everyone had an equal opportunity regardless of personal characteristics:

> *I think that fairness is that it's inclusive, everybody has the opportunity, regardless*
>
> *(Dave, People lead)*

> *Just that I am considering everybody equally and I'm considering the same things about each person. So that I'm not taking into account, you know, like I said, the age and the race and stuff like that shouldn't matter. I'm looking at the important things with each person. I'm looking at work history. I'm looking at you know, I'm looking at the stuff that matters for 'are you going to come here and be a good worker?' … Like, 'are you, are you going to show up?'. Like, I'm looking at the indicators of that.*
>
> *(Lindsey, People lead)*

However, the project team were conscious that whatever technology was used in hiring decisions the throughput and outcomes would need to be checked for disparate outcomes. Checks were therefore built into algorithm development as "one of the critical steps to the validation" and monitored on an ongoing basis using monitoring data collected from applicants but separated out from applications.

> *We feel, like, moving in this direction can clear up that bias and get us on a path of consistency. Right. And as long as we're monitoring and evaluating those algorithms and those models, we can make them smarter.*
>
> *(Project lead)*

As noted above, use of a Fair Aware Classifier algorithm was aimed at increasing precision in predicting outcomes, while at the same time removing variance in the data that is related to protected characteristics. This was hoped to minimize any potential bias related to protected characteristics, thus minimising the potential for disparate outcomes, while at the same time avoiding potential disparate treatment by predicting the best person for the job. However, the team were cognizant that in data science achieving individual fairness and group fairness were not always compatible, and that there could be a tension between the two.

> *Demographic parity is still widely being considered by the regulatory authorities. Also, as you said we wanted the best people to get the job. Meaning you still are looking for the model being accurate in predicting their performance, right? We want it to be equally accurate across the different*

> *demographic subgroups. And these two things, most of the time, will not, mathematically, will not accomplish that.*
>
> *(Technical lead)*

While users did not report needing to monitor hiring figures to make sure hires reflected the broader population where they were located, all felt that their associates were at least as diverse as the population in their wider market. Thus, group fairness was felt to result from applying individual fairness.

Role of the Hiring Helper in reducing bias and making hiring fair

Making greater use of the Hiring Helper was generally seen by users as potentially removing or at least reducing human bias from hiring decisions. One user felt that the developments to the Hiring Helper had made hiring more data-driven and therefore less biased. While he felt the algorithm was not better than a human manager at predicting good hires, it was as good as most managers and better than less experienced managers. The manager viewed ADM as unproblematic, in the sense of removing bias, and did not feel any additional adjustments were needed to ensure that hiring decisions were 'fair' or unbiased:

> *I think exactly what we're looking at is the biggest step, we've made in a long time. I mean, I think taking a lot of those things out of the equation. You have no preconceived idea about this person when they walk in the door because you don't know anything about 'em. So you've basically created a system that helps to eliminate bias, more so than anything I've seen in a long time. … So I think it's going to give us a better, maybe a more diverse group of initial candidates to get into the company.*
>
> *(Joe, Store manager)*

There was recognition, though, among the project team and people leads who were aware of the technology involved, that AI was not entirely free from the potential for bias. However, the risk was often assumed to be due to poor design choices or that the algorithm might reflect the data engineer's own potential biases. Data engineers who are white and male, for example, might not recognise or pick up on biases in the system.

> *Ultimately, somebody has to initially build that algorithm, you know, train the AI and put all of those inputs and variables into the equation and that's ultimately going to be where the genesis of bias in machine learning and AI could show up. If we're not intentional and thoughtful, you know, in trying to recognise and mitigate bias in that process, then it's going to get engineered into the AI.*
>
> *(Project team respondent)*

A number of researchers have identified bias in historical hiring decisions used in training data as a potential source of bias in algorithms used in shortlisting (Kamishima et al., 2012; Sánchez-Mondero et al., 2020). This was not cited as a potential source of bias among store-level respondents or those on the project team. On one hand, as the algorithm was trained to predict 90-day turnover from a hiring cohort, and not historical hiring decisions, any bias in hiring decisions is unlikely to have been encoded into the algorithm. On the other hand, if there is systematic bias in 90-day turnover rates among different groups there is a risk that this bias could be reflected in the training data if not identified and corrected for in the development of the algorithm. It is this type of bias that the fair aware classifier is hoped to eliminate.

> *We have been researching and following the trend of Fair AI from back in 2017. [With a Fair Aware Classifier] you would be able to achieve a business outcome with a fair outcome too. … So that's one of the advantages. You are able to achieve a consistent better outcome by utilizing the algorithm. And also it's still a fair outcome. The beauty of this algorithm [is that it helps] to strip out the information that are overly sensitive to gender and ethnicity, but still retain the useful information that helps it with a prediction.*
>
> *(Project Team developer)*

Even with a fair aware classifier potential issues related to training data can arise. Achieving an accurate and fair algorithm is dependent upon having a large amount of data in the training data, something that is unlikely to be an issue in this case, and it is important to ensure that the training data reflects the target population in terms of individual characteristics.

While this was something the development team looked at, potential bias could emerge if there is a systematic difference between the training data and the target population which is found to be related to a protected characteristic. ‘Population bias’ and ‘longitudinal data fallacy’ would be examples of this type of issue (Mehrabi et al. 2019). For example, if there is a systematic difference between applicants who were hired and those who were not hired this might mean that the algorithm is good at predicting turnover rates for those who were hired but may not be as good at predicting the turnover rates of those who were not hired.

Thus, while the algorithm would be accurate and unbiased when predicting turnover rates in the training data it may be less accurate and unbiased when applied to all applicants. A similar issue might arise if the training cohort was systematically different from the current application pool for some reason (e.g. temporary unemployment due to the Covid). These issues were not raised specifically during interviews with the project team but could be something to look into in future.

Role of removing the in-person interview on making hiring fair and less biased

Removing the in-person interview was generally seen as a positive step in terms of minimising potential human bias. Interviewers would not necessarily know the age or ethnicity of candidates, for example, and would not be influenced by preconceived ideas related to appearance.

> *I think that it could actually be considered more fair because people can't really see the person and make a judgment from their disposition or race or colour or age. I mean you can guess over the phone, but I feel as if you're going to go by exactly the application, their communication, what they're verbally saying to you, and if it's fitting your needs.*
>
> *(Kat, People lead)*

However, two of the people leads interviewed noted that telephone interviews might not suit all candidates and might favour those who were more confident over the phone.

> *You know, sometimes your body language may give us something different, you know, your voice, your tone as well. So it could work against you. It just depends on if you are a great communicator over the phone or if you're easier to understand in-person off of body language or physicality.*
>
> *(Kat, People lead)*

This was recognised as a potential issue by the project team, who also considered other interview formats, such as video interviews. But they felt that these were prone to additional potential bias related to the appearance of the candidate’s home as well as issues around access to technology. As a consequence telephone interviews were considered least prone to such issues.

Respondents gave examples of where human bias could surface in in-person interviews. For example, hiring managers might raise concerns about an applicant’s ability to do a given job, either due to age, pregnancy or disability and people leads have directed them to not take such factors into account. One of the project team cited an example of a hiring manager asking the store manager whether a candidate, who was a single amputee and had passed initial screening in-person, should take part in a physical work assessment because they were concerned that they might not be able to complete it. The candidate was allowed to do the assessment, performed well, and was ultimately hired. The respondent commented:

> *But had that bias not been checked by that more senior manager that individual might not have been brought back for the assessment, not given the opportunity to show what they were able to do and invariably not gotten the job. You know, if you eliminate that that face-to-face, or that visual first step interaction, and that's on the phone, that probably never happens.*
>
> *(Project team respondent)*

Similarly, one people lead reported hiring an older candidate for an online grocery pick-up job, who they might have assumed could not do such a physical job if they had interviewed her in-person, but had turned out to be a good associate. And the store manager interviewed gave a similar example.

> *I don't know that, if we would have interviewed her to her face, we would have been confident that she can handle online grocery pickup. And so anyway, we get her in here for the orientation and I was kind of worried, but she runs circles around those younger kids and does a great job. So I think sometimes the face-to-face. I mean, like I said, I don't know that we would have picked her for OGP [online grocery pick-up] just based on her appearance but on the phone I knew she was older and I took that chance and it worked out.*
>
> *(Maxine, People lead)*

These examples show that removing the in-person interview can remove potential human bias related to appearance, but they also show that human decision makers can be reflexive. They can recognise situations, ask questions and self-correct. While ADM can make data-driven predictions, if there is bias in the algorithm or the training data the computer is unable to reflect on this itself. It would be up to the ML engineer to recognise and identify the potential problems. While the project team showed that they were aware of the potential challenges and made efforts to address them, careful self-reflection is warranted because in some cases it can be difficult to identify coding issues, if missed first time round.

Impact of awareness and trust on the objective of making hiring fairer

While not all users were aware that changes made to the hiring system were intended to reduce human bias most felt that removing the in-person interview helped minimise potential human bias as well as making the process safer and faster during the pandemic. Where it was understood that the algorithm used to sort candidates in the Hiring Helper aimed to predict employment outcomes, this was generally felt to make the hiring process more data-driven and so fairer.

However, lack of confidence in the pre-employment assessment and the algorithm's ability to reliably predict quality candidates threatens to undermine this objective. At least one respondent reported reverting to in-person interviews and others were not using the list as intended. This potentially opens the process up to the influence of other factors unrelated to outcomes such as appearance, confidence in telephone calls or proximity to the store, some of which could potentially be correlated to protected characteristics.

While the algorithm and pre-employment assessment were intended to reduce potential bias, its ability to do so is not a given and will depend upon an ongoing process of development and auditing.

In order to ensure that the changes to the hiring system are successful in ensuring hiring is fair and unbiased it will be important to:

1) Build trust in the Hiring Helper by increasing awareness and ensuring recommendations are reliable;
2) Continue to monitor and assess the algorithm and its output to ensure that there is no bias in the system;
3) Continue efforts to identify and address potential systematic sources of bias in other parts of the system (e.g. that telephone interviews do not systematically favour some groups of candidates over others)

5. Discussion

Advances in AI and predictive analytics have greatly expanded the range of tasks that computers can perform, including its application to all stages of the hiring process (Raghavan et al., 2020). Computers using AI have the potential to make decisions and predictions quicker and more effectively than humans (Upadhyay and Khandelwal, 2018) and to remove human bias (Houser, 2019; Cowgill, 2018; Bogen and Reike, 2018). Increased demand for store-level hourly paid associates due to the effects of the pandemic accelerated the need for a hiring system that is safe, fast, effective and fair. Walmart therefore sought to leverage the potential of AI in the hiring system for store-level hourly paid associates. The reported objectives of the Rapid Recruitment project were to:

1) Improve hiring decisions (in terms of 90-day turnover rates);
2) Make hiring fairer and less prone to bias;
3) Speed up the hiring process; and
4) Protect the safety of associates and applicants during the pandemic.

Walmart implemented a number of changes to the hiring system in order to achieve this, including:

- Development of an algorithm (using machine learning) to predict 90-day turnover;
- Removing in-person interviews;
- Enabling people leads and store managers to make job offers over the phone;
- Advising people leads and store managers to make greater use of list position in the Hiring Helper to guide shortlisting decisions; and
- Changes to the Hiring Helper UI to make candidates at the top of the list and reasons for their position more prominent.

Users (people leads and store managers) interviewed had implemented the above changes and were, by and large, making greater use of the Hiring Helper. However, the extent to which users were making use of the Hiring Helper to guide hiring decisions varied according to the level of awareness they had about the technology involved and/or confidence in the system's ability to make reliable recommendations. This lack of trust and awareness translated into behaviours that potentially undermine some of the objectives of the changes.

Protecting health and safety and speeding up the hiring process

All of those interviewed reported that the recent changes had indeed sped up the hiring process. They all felt that removing the in-person interview was a necessary step to protect associates and candidates during the pandemic. These outcomes were universally valued by respondents and seen as an advantage.

However, lack of awareness about how candidates were sorted in the Hiring Helper and lack of confidence in the sorting led to one or two users were reviewing more applications than was perhaps necessary. Furthermore, one respondent felt that replacing the in-person interview with telephone interviews weakened their ability to gauge important qualities, such as attitude and professionalism, through body language and appearance. This had led some hiring managers to revert to in-person interviews, slowing down the process and increasing the risk of Covid transmission.

Improving hiring outcomes

One of the main objectives of the changes was to improve hiring outcomes by making hiring decisions data-driven and minimizing human bias related to appearance and other factors unrelated to performance. In order to achieve this, first, the algorithm used in the Hiring Helper must be at least as good as human hirers at predicting the desired outcomes, and second, users need to make use of those predictions.

The project team reported that piloting and monitoring of the system indicated that the algorithm does indeed outperform human hirers at predicting 90-day turnover. While this is hard to confirm without having access to the data, some users felt that 90-day turnover rates had improved but that the pandemic made this hard to assess.

However, some users had reservations about the pre-employment assessment and the algorithm's ability to predict 'good' hires. Some reported finding candidates at the top of the list who they felt would not make good associates or found candidates rated as 'poor' who they felt would make good associates. This could at least in part be due to a discrepancy between the outcomes that the algorithm and hiring managers are trying to predict.

The algorithm attempts to predict 90-day turnover, whereas people leads and hiring managers may be trying to predict other characteristics such as customer service and potential performance. While some candidates might perform well and/or have good customer service skills, other factors may make them less likely to stay in post for more than 90 days. Thus, lack of awareness that the algorithm is trying to predict turnover and not some other performance measure might undermine confidence in the sorting used in the Hiring Helper and make users less likely to use the list as intended.

Making hiring fairer and less open to bias

Diversity and inclusion are important goals for the organization and are not just understood to be a legal and moral imperative but are also felt to help to support innovation and productivity. The use of predictive analytics and replacing in-person interviews with shorter telephone interviews aimed to minimise potential human bias and make hiring fairer.

How successful the changes are in achieving this objective is largely dependent upon two challenges:

1) Making sure that the algorithm is accurate and unbiased;
2) Making sure user behaviour does not lead to biased decisions.

In relation to the first challenge, using a Fair Aware Classifier aimed to predict 90-day turnover while removing direct and indirect bias that can arise when measures used in the prediction model are correlated with protected characteristics (e.g. age, ethnicity and gender). However, other potential sources of bias can arise from the training data.

Firstly, bias could arise if there is a systematic bias in 90-day turnover rates in the training data (a cohort of historical hires) that is not fully corrected for by the fair aware classifier. Secondly, bias can arise if the training data is systematically different from the population in some way. 'Population bias' could result if the relationship between predictive measures used in the algorithm and 90-day turnover rates is different for non-hires compared to hires. The ‘longitudinal data fallacy’ would apply if the current cohort of applicants is systematically different from the training cohort. In either situation the algorithm may be accurate and unbiased in the training data but less accurate and unbiased when applied to the target population.

In relation to the second challenge, most respondents felt that removing the in-person interview removed the potential for human bias related to appearance and other characteristics. However, some users raised reservations about the loss of the ‘human touch’, ability to assess candidates effectively over the phone and that telephone interviews may favour some candidates over others. Again, lack of awareness and trust in the system could undermine the objectives of the changes if it means users bypass recommendations or revert to in-person interviews (assuming these are more prone to human bias).

6. Conclusion and Recommendations

The Rapid Recruitment project undoubtedly reduced the risk of Covid transmission and sped up recruitment, enabling the hiring of more than 460,000 associates during the pandemic. Understanding of what the Hiring Helper is attempting to predict and trust in its ability to do so is an important factor in achieving the project's stated objectives. This was acknowledged by the project team who saw building trust in the system as an ongoing journey involving listening and responding to user concerns. The project team also recognised that accuracy and transparency were important factors in this. Portability, ongoing development of the algorithm and introduction of the 'explain to me' function can be seen as attempts to address this.

However, some users were not fully aware of how candidates were sorted in the Hiring Helper. Increasing awareness of the technology and the target measure used in the sorting may help users understand any misalignment between candidates' position on the list and users' own assessment of suitability. This would help build trust in the system and enable users to better use of the Hiring Helper in hiring decisions. They could then balance the likelihood of associates staying in post alongside other factors such as potential performance and customer service skills.

As noted by the project team, ensuring the algorithm is accurate and unbiased is part of the process of building trust in the system. In order to achieve this it will be important to:

1) Iron out any inaccuracies in the algorithm and business rules;
2) Ensure training data and cohort reflects the target population and is not systematically different in some way;
3) Ensure that any bias in 90-day turnover rates among historical hires is fully controlled for in the algorithm;
4) Continue to monitor hiring decisions for disparate outcomes on an ongoing basis.

References

Albert, E. T. (2019). AI in talent acquisition: a review of AI-applications used in recruitment and selection. *Strategic HR Review*, 18(5), 215-221.

Allen, R. & Masters, D. (2020). Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. *ERA Forum*, 20, 585–598. https://doi.org/10.1007/s12027-019-00582-w

Bogen, M. & Rieke, A. (2018). *Help Wanted: An Exploration of Hiring Algorithms, Equity and Bias*. www.upturn.org/hiring-algorithms

Booth, R. (2019, October 25). Unilever saves on recruiters by using AI to assess job interviews. www.TheGuardian.com

Brione, P. (2020). *My Boss the Algorithm: An ethical look at algorithms in the workplace*. London: Advisory, Conciliation and Arbitration Service.

Brynjolfsson, E., Rock, D. & Syverson, C. (2017). *Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics*. NBER Working Paper 24001.

Chen, L., Ma, R., Hannák, A. & Wilson, C., (2018). *Investigating the impact of gender on rank in resume search engines*. In Proceedings of the 2018 chi conference on human factors in computing systems (pp. 1-14).

Cowgill, B. (2018). *Bias and Productivity in Humans and Algorithms: Theory and evidence from resume screening*. Columbia Business School, Columbia University, 29. pp 1-35.

D'Alessandro, B., O'Neil, C. & La Gatta, T. (2017). Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data*, 5(2), 120-134. DOI: 10.1089/big.2016.0048

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. www.Reuters.com.

Ernst, E., Merola, R. and Samaan, D. (2018). *The Economics of Artificial Intelligence: Implications for the Future of Work*. ILO Future of Work Research Paper Series. Geneva: International Labour Organization.

Eubanks, V. (2018). *Automating Inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Frey, C. & Osborne, M. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, pp.254-280.

Heric, M. (2018, October 10). HR's New Digital Mandate. Bain and Company brief. www.bain.com

Hopping, C. (2015, June 5). The truth about talent selection algorithms. www.launchpadrecruits.com.

Houser, K.A., 2019. Can AI Solve the Diversity Problem in the Tech Industry: Mitigating Noise and Bias in Employment Decision-Making. *Stanford Technology Law Review*, 22, 290-353.

Hunt, V., Layton, D. & Prince, S. (2015). *Diversity Matters*. London: Mc Kinsey and Company.

Kamishima, T., Akaho, S., Asoh, H. & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regulizer. In P. A. Flach, T. De Bie & N. Cristianini (eds), *Machine Learning and Knowledge Discovery in Databases*. ECML PKDD 2012. Lecture Notes in Computer Science, vol 7524. Berlin: Springer.

Lambrecht, A. & Tucker, C. E. (2019). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7), 2966-2981.

Lindebaum, D., Vesa, M. and den Hond, F., (2020). Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review*, 45(1), pp.247-263. https://doi.org/10.5465/amr.2018.0181

Marr, B. (2018, December 14). The Amazing Ways How Unilever Uses Artificial Intelligence To Recruit & Train Thousands Of Employees. www.Forbes.com

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A., (2019). *A survey on bias and fairness in machine learning*. arXiv preprint arXiv:1908.09635.

Newman, D., Fast, N. & Harmon, D. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160, 149-167.

O'Neil, C. (2016). *Weapons of Math Destruction: How big data increases inequality and threatens democracy*. New York: Random House.

Pessach, D., Singer, G., Avrahami, D., Chalutz Ben-Gal, H., Shmueli, E. & Ben-Gal, I. (2020). Employees Recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134, 1-18.

Raghavan, M., Barocas, S., Kleinberg, J. & Levy, K., (2020). *Mitigating bias in algorithmic hiring: Evaluating claims and practices*. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469-481).

Sánchez-Mondero, 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*. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27-30, 2020, Barcelona, Spain, http://dx.doi.org/10.2139/ssrn.3463141

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

Upadhyay, A. & Khandelwal, K. (2018). Applying Artificial Intelligence: implications for recruitment. *Strategic HR Review*, 17(5), 255-258.

Van Esch, P. & Black, J. (2019). Factors that Influence New Generation Candidates to Engage with and Complete Digital, AI-enabled Recruiting. *Business Horizons*, 62(6), pp.729-739.